

COMS 4771 Lecture 16

1. Fixed-design linear regression
2. Ridge and principal components regression
3. Sparse regression and Lasso

FIXED-DESIGN LINEAR REGRESSION

FIXED-DESIGN LINEAR REGRESSION

A simplified fixed-design setting

$\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^p$ assumed to be *fixed*—i.e., not random;
 $y^{(1)}, y^{(2)}, \dots, y^{(n)}$ are independent random variables, with

$$\mathbb{E}(y^{(i)}) = \langle \mathbf{x}^{(i)}, \mathbf{w}_\star \rangle, \quad \text{var}(y^{(i)}) = \sigma^2.$$

FIXED-DESIGN LINEAR REGRESSION

A simplified fixed-design setting

$\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^p$ assumed to be *fixed*—i.e., not random;
 $y^{(1)}, y^{(2)}, \dots, y^{(n)}$ are independent random variables, with

$$\mathbb{E}(y^{(i)}) = \langle \mathbf{x}^{(i)}, \mathbf{w}_\star \rangle, \quad \text{var}(y^{(i)}) = \sigma^2.$$

$$\underbrace{\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}}_{\mathbf{y} \in \mathbb{R}^n} = \underbrace{\begin{bmatrix} - & \mathbf{x}^{(1)\top} & - \\ - & \mathbf{x}^{(2)\top} & - \\ & \vdots & \\ - & \mathbf{x}^{(n)\top} & - \end{bmatrix}}_{\mathbf{X} \in \mathbb{R}^{n \times p}} \underbrace{\begin{bmatrix} w_{\star,1} \\ w_{\star,2} \\ \vdots \\ w_{\star,p} \end{bmatrix}}_{\mathbf{w}_\star \in \mathbb{R}^p} + \underbrace{\begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(n)} \end{bmatrix}}_{\boldsymbol{\varepsilon} \in \mathbb{R}^n}$$

where $\varepsilon^{(1)}, \varepsilon^{(2)}, \dots, \varepsilon^{(n)}$ are independent, $\mathbb{E}(\varepsilon^{(i)}) = 0$, and $\text{var}(\varepsilon^{(i)}) = \sigma^2$.

FIXED-DESIGN LINEAR REGRESSION

A simplified fixed-design setting

$\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^p$ assumed to be *fixed*—i.e., not random;
 $y^{(1)}, y^{(2)}, \dots, y^{(n)}$ are independent random variables, with

$$\mathbb{E}(y^{(i)}) = \langle \mathbf{x}^{(i)}, \mathbf{w}_\star \rangle, \quad \text{var}(y^{(i)}) = \sigma^2.$$

$$\underbrace{\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}}_{\mathbf{y} \in \mathbb{R}^n} = \underbrace{\begin{bmatrix} - & \mathbf{x}^{(1)\top} & - \\ - & \mathbf{x}^{(2)\top} & - \\ & \vdots & \\ - & \mathbf{x}^{(n)\top} & - \end{bmatrix}}_{\mathbf{X} \in \mathbb{R}^{n \times p}} \underbrace{\begin{bmatrix} w_{\star,1} \\ w_{\star,2} \\ \vdots \\ w_{\star,p} \end{bmatrix}}_{\mathbf{w}_\star \in \mathbb{R}^p} + \underbrace{\begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(n)} \end{bmatrix}}_{\boldsymbol{\varepsilon} \in \mathbb{R}^n}$$

where $\varepsilon^{(1)}, \varepsilon^{(2)}, \dots, \varepsilon^{(n)}$ are independent, $\mathbb{E}(\varepsilon^{(i)}) = 0$, and $\text{var}(\varepsilon^{(i)}) = \sigma^2$.

Want to find $\hat{\mathbf{w}} \in \mathbb{R}^p$ based on $(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(n)}, y^{(n)})$ so that

$$\mathbb{E} \left[\frac{1}{n} \|\mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}\|_2^2 \right]$$

is small (e.g., $o(1)$ as function of n).

ORDINARY LEAST SQUARES

Recall: assuming $\mathbf{X}^\top \mathbf{X}$ is invertible,

$$\hat{\mathbf{w}}_{\text{ols}} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

ORDINARY LEAST SQUARES

Recall: assuming $\mathbf{X}^\top \mathbf{X}$ is invertible,

$$\hat{\mathbf{w}}_{\text{ols}} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Expressing $\mathbf{X}\hat{\mathbf{w}}_{\text{ols}}$ in terms of $\mathbf{X}\mathbf{w}_\star$:

$$\mathbf{X}\hat{\mathbf{w}}_{\text{ols}}$$

ORDINARY LEAST SQUARES

Recall: assuming $\mathbf{X}^\top \mathbf{X}$ is invertible,

$$\hat{\mathbf{w}}_{\text{ols}} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Expressing $\mathbf{X}\hat{\mathbf{w}}_{\text{ols}}$ in terms of $\mathbf{X}\mathbf{w}_\star$:

$$\mathbf{X}\hat{\mathbf{w}}_{\text{ols}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\mathbf{w}_\star + \boldsymbol{\varepsilon})$$

ORDINARY LEAST SQUARES

Recall: assuming $\mathbf{X}^\top \mathbf{X}$ is invertible,

$$\hat{\mathbf{w}}_{\text{ols}} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Expressing $\mathbf{X}\hat{\mathbf{w}}_{\text{ols}}$ in terms of $\mathbf{X}\mathbf{w}_\star$:

$$\begin{aligned} \mathbf{X}\hat{\mathbf{w}}_{\text{ols}} &= \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\mathbf{w}_\star + \boldsymbol{\varepsilon}) \\ &= \mathbf{X}\mathbf{w}_\star + \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon} \end{aligned}$$

ORDINARY LEAST SQUARES

Recall: assuming $\mathbf{X}^\top \mathbf{X}$ is invertible,

$$\hat{\mathbf{w}}_{\text{ols}} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Expressing $\mathbf{X}\hat{\mathbf{w}}_{\text{ols}}$ in terms of $\mathbf{X}\mathbf{w}_\star$:

$$\begin{aligned} \mathbf{X}\hat{\mathbf{w}}_{\text{ols}} &= \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\mathbf{w}_\star + \boldsymbol{\varepsilon}) \\ &= \mathbf{X}\mathbf{w}_\star + \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon} \\ &= \mathbf{X}\mathbf{w}_\star + \boldsymbol{\Pi}\boldsymbol{\varepsilon} \end{aligned}$$

where $\boldsymbol{\Pi} \in \mathbb{R}^{n \times n}$ is the orthogonal projection operator for $\text{ran}(\mathbf{X})$.

ORDINARY LEAST SQUARES

Recall: assuming $\mathbf{X}^\top \mathbf{X}$ is invertible,

$$\hat{\mathbf{w}}_{\text{ols}} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Expressing $\mathbf{X}\hat{\mathbf{w}}_{\text{ols}}$ in terms of $\mathbf{X}\mathbf{w}_\star$:

$$\begin{aligned} \mathbf{X}\hat{\mathbf{w}}_{\text{ols}} &= \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\mathbf{w}_\star + \boldsymbol{\varepsilon}) \\ &= \mathbf{X}\mathbf{w}_\star + \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon} \\ &= \mathbf{X}\mathbf{w}_\star + \boldsymbol{\Pi}\boldsymbol{\varepsilon} \end{aligned}$$

where $\boldsymbol{\Pi} \in \mathbb{R}^{n \times n}$ is the orthogonal projection operator for $\text{ran}(\mathbf{X})$.

$$\mathbb{E} \left[\frac{1}{n} \|\mathbf{X}\mathbf{w}_\star - \mathbf{X}\hat{\mathbf{w}}_{\text{ols}}\|_2^2 \right] = \frac{\sigma^2 p}{n}.$$

RIDGE REGRESSION AND PRINCIPAL COMPONENTS REGRESSION

RIDGE REGRESSION

- ▶ Ordinary least squares only applies if $\mathbf{X}^\top \mathbf{X}$ is invertible, which is not possible if $p > n$.

RIDGE REGRESSION

- ▶ Ordinary least squares only applies if $\mathbf{X}^\top \mathbf{X}$ is invertible, which is not possible if $p > n$.
- ▶ **Ridge regression** (Hoerl, 1962): for $\lambda > 0$,

$$\hat{\mathbf{w}}_\lambda := \arg \min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2.$$

RIDGE REGRESSION

- ▶ Ordinary least squares only applies if $\mathbf{X}^\top \mathbf{X}$ is invertible, which is not possible if $p > n$.
- ▶ **Ridge regression** (Hoerl, 1962): for $\lambda > 0$,

$$\hat{\mathbf{w}}_\lambda := \arg \min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2.$$

- ▶ Gradient of objective is zero when

$$(\mathbf{X}^\top \mathbf{X} + n\lambda \mathbf{I})\mathbf{w} = \mathbf{X}^\top \mathbf{y},$$

which *always* has a unique solution (since $\lambda > 0$):

$$\hat{\mathbf{w}}_\lambda := (\mathbf{X}^\top \mathbf{X} + n\lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}.$$

RIDGE REGRESSION

- ▶ Ordinary least squares only applies if $\mathbf{X}^\top \mathbf{X}$ is invertible, which is not possible if $p > n$.
- ▶ **Ridge regression** (Hoerl, 1962): for $\lambda > 0$,

$$\hat{\mathbf{w}}_\lambda := \arg \min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2.$$

- ▶ Gradient of objective is zero when

$$(\mathbf{X}^\top \mathbf{X} + n\lambda \mathbf{I})\mathbf{w} = \mathbf{X}^\top \mathbf{y},$$

which *always* has a unique solution (since $\lambda > 0$):

$$\hat{\mathbf{w}}_\lambda := (\mathbf{X}^\top \mathbf{X} + n\lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}.$$

RIDGE REGRESSION

- ▶ Ordinary least squares only applies if $\mathbf{X}^\top \mathbf{X}$ is invertible, which is not possible if $p > n$.
- ▶ **Ridge regression** (Hoerl, 1962): for $\lambda > 0$,

$$\hat{\mathbf{w}}_\lambda := \arg \min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2.$$

- ▶ Gradient of objective is zero when

$$(\mathbf{X}^\top \mathbf{X} + n\lambda \mathbf{I})\mathbf{w} = \mathbf{X}^\top \mathbf{y},$$

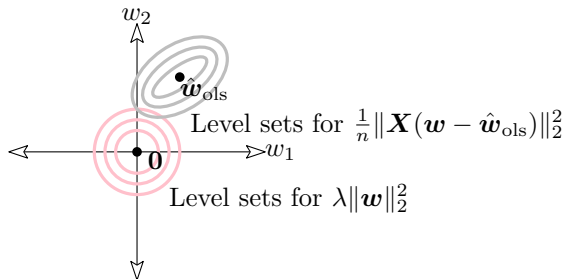
which *always* has a unique solution (since $\lambda > 0$):

$$\hat{\mathbf{w}}_\lambda := \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{y} \right).$$

RIDGE REGRESSION: GEOMETRY

Ridge regression objective (as function of $\mathbf{w}$) can be written as

$$\frac{1}{n} \|\mathbf{X}(\mathbf{w} - \hat{\mathbf{w}}_{\text{ols}})\|_2^2 + \lambda \|\mathbf{w}\|_2^2 + (\text{stuff not depending on } \mathbf{w})$$



RIDGE REGRESSION: EIGENDECOMPOSITION

Write eigendecomposition of $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ as

$$\frac{1}{n}\mathbf{X}^\top\mathbf{X} = \sum_{j=1}^p \lambda_j \mathbf{v}_j \mathbf{v}_j^\top$$

where $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p \in \mathbb{R}^p$ are orthonormal eigenvectors with corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$.

RIDGE REGRESSION: EIGENDECOMPOSITION

Write eigendecomposition of $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ as

$$\frac{1}{n}\mathbf{X}^\top\mathbf{X} = \sum_{j=1}^p \lambda_j \mathbf{v}_j \mathbf{v}_j^\top$$

where $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p \in \mathbb{R}^p$ are orthonormal eigenvectors with corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$.

- Eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$ comprise an *orthonormal basis* for $\mathbb{R}^p$.

We'll look at $\hat{\mathbf{w}}_{\text{ols}}$ and $\mathbf{w}_\star$ in this basis: e.g.,

$$\hat{\mathbf{w}}_{\text{ols}} = \sum_{j=1}^p \langle \mathbf{v}_j, \hat{\mathbf{w}}_{\text{ols}} \rangle \mathbf{v}_j.$$

RIDGE REGRESSION: EIGENDECOMPOSITION

Write eigendecomposition of $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ as

$$\frac{1}{n}\mathbf{X}^\top\mathbf{X} = \sum_{j=1}^p \lambda_j \mathbf{v}_j \mathbf{v}_j^\top$$

where $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p \in \mathbb{R}^p$ are orthonormal eigenvectors with corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$.

- Eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$ comprise an *orthonormal basis* for $\mathbb{R}^p$.

We'll look at $\hat{\mathbf{w}}_{\text{ols}}$ and $\mathbf{w}_\star$ in this basis: e.g.,

$$\hat{\mathbf{w}}_{\text{ols}} = \sum_{j=1}^p \langle \mathbf{v}_j, \hat{\mathbf{w}}_{\text{ols}} \rangle \mathbf{v}_j.$$

- The inverse of $\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ has the form

$$\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I} \right)^{-1} = \sum_{j=1}^p \frac{1}{\lambda_j + \lambda} \mathbf{v}_j \mathbf{v}_j^\top.$$

RIDGE REGRESSION VS. ORDINARY LEAST SQUARES

If $\hat{\mathbf{w}}_{\text{ols}}$ exists, then

RIDGE REGRESSION VS. ORDINARY LEAST SQUARES

If $\hat{\mathbf{w}}_{\text{ols}}$ exists, then

$$\hat{\mathbf{w}}_{\lambda} = \left(\frac{1}{n} \mathbf{X}^{\top} \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \left(\frac{1}{n} \mathbf{X}^{\top} \mathbf{X} \right) \hat{\mathbf{w}}_{\text{ols}}$$

RIDGE REGRESSION VS. ORDINARY LEAST SQUARES

If $\hat{\mathbf{w}}_{\text{ols}}$ exists, then

$$\begin{aligned}\hat{\mathbf{w}}_{\lambda} &= \left(\frac{1}{n} \mathbf{X}^{\top} \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \left(\frac{1}{n} \mathbf{X}^{\top} \mathbf{X} \right) \hat{\mathbf{w}}_{\text{ols}} \\ &= \left(\sum_{j=1}^p \frac{1}{\lambda_j + \lambda} \mathbf{v}_j \mathbf{v}_j^{\top} \right) \left(\sum_{j=1}^p \lambda_j \mathbf{v}_j \mathbf{v}_j^{\top} \right) \hat{\mathbf{w}}_{\text{ols}}\end{aligned}$$

RIDGE REGRESSION VS. ORDINARY LEAST SQUARES

If $\hat{\mathbf{w}}_{\text{ols}}$ exists, then

$$\begin{aligned}\hat{\mathbf{w}}_{\lambda} &= \left(\frac{1}{n} \mathbf{X}^{\top} \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \left(\frac{1}{n} \mathbf{X}^{\top} \mathbf{X} \right) \hat{\mathbf{w}}_{\text{ols}} \\ &= \left(\sum_{j=1}^p \frac{1}{\lambda_j + \lambda} \mathbf{v}_j \mathbf{v}_j^{\top} \right) \left(\sum_{j=1}^p \lambda_j \mathbf{v}_j \mathbf{v}_j^{\top} \right) \hat{\mathbf{w}}_{\text{ols}} \\ &= \left(\sum_{j=1}^p \frac{\lambda_j}{\lambda_j + \lambda} \mathbf{v}_j \mathbf{v}_j^{\top} \right) \hat{\mathbf{w}}_{\text{ols}} \quad (\text{by orthogonality})\end{aligned}$$

RIDGE REGRESSION VS. ORDINARY LEAST SQUARES

If $\hat{\mathbf{w}}_{\text{ols}}$ exists, then

$$\begin{aligned}\hat{\mathbf{w}}_{\lambda} &= \left(\frac{1}{n} \mathbf{X}^{\top} \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \left(\frac{1}{n} \mathbf{X}^{\top} \mathbf{X} \right) \hat{\mathbf{w}}_{\text{ols}} \\&= \left(\sum_{j=1}^p \frac{1}{\lambda_j + \lambda} \mathbf{v}_j \mathbf{v}_j^{\top} \right) \left(\sum_{j=1}^p \lambda_j \mathbf{v}_j \mathbf{v}_j^{\top} \right) \hat{\mathbf{w}}_{\text{ols}} \\&= \left(\sum_{j=1}^p \frac{\lambda_j}{\lambda_j + \lambda} \mathbf{v}_j \mathbf{v}_j^{\top} \right) \hat{\mathbf{w}}_{\text{ols}} \quad (\text{by orthogonality}) \\&= \sum_{j=1}^p \frac{\lambda_j}{\lambda_j + \lambda} \langle \mathbf{v}_j, \hat{\mathbf{w}}_{\text{ols}} \rangle \mathbf{v}_j.\end{aligned}$$

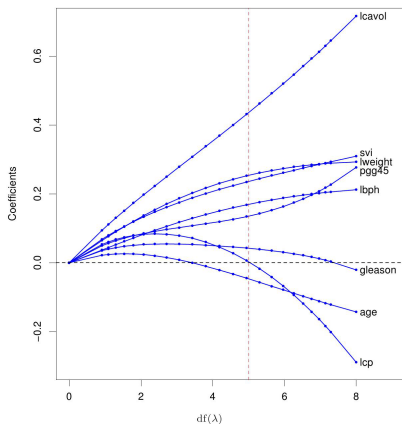
RIDGE REGRESSION VS. ORDINARY LEAST SQUARES

If $\hat{\mathbf{w}}_{\text{ols}}$ exists, then

$$\begin{aligned}\hat{\mathbf{w}}_{\lambda} &= \left(\frac{1}{n} \mathbf{X}^{\top} \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \left(\frac{1}{n} \mathbf{X}^{\top} \mathbf{X} \right) \hat{\mathbf{w}}_{\text{ols}} \\&= \left(\sum_{j=1}^p \frac{1}{\lambda_j + \lambda} \mathbf{v}_j \mathbf{v}_j^{\top} \right) \left(\sum_{j=1}^p \lambda_j \mathbf{v}_j \mathbf{v}_j^{\top} \right) \hat{\mathbf{w}}_{\text{ols}} \\&= \left(\sum_{j=1}^p \frac{\lambda_j}{\lambda_j + \lambda} \mathbf{v}_j \mathbf{v}_j^{\top} \right) \hat{\mathbf{w}}_{\text{ols}} \quad (\text{by orthogonality}) \\&= \sum_{j=1}^p \frac{\lambda_j}{\lambda_j + \lambda} \langle \mathbf{v}_j, \hat{\mathbf{w}}_{\text{ols}} \rangle \mathbf{v}_j.\end{aligned}$$

Interpretation: Shrink $\hat{\mathbf{w}}_{\text{ols}}$ towards zero by $\frac{\lambda_j}{\lambda_j + \lambda}$ factor in direction $\mathbf{v}_j$.

COEFFICIENT PROFILE



Horizontal axis: varying λ (large λ to left, small λ to right).

Vertical axis: coefficient value in $\hat{w}_\lambda$ for eight different variables.

RIDGE REGRESSION: THEORY

Theorem:

$$\mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_\lambda \|_2^2 \right] = \underbrace{\lambda \sum_{j=1}^p \frac{\lambda_j \lambda}{(\lambda_j + \lambda)^2} \langle \mathbf{v}_j, \mathbf{w}_\star \rangle^2}_{\text{bias}} + \underbrace{\frac{\sigma^2}{n} \sum_{j=1}^p \left(\frac{\lambda_j}{\lambda_j + \lambda} \right)^2}_{\text{variance}}.$$

RIDGE REGRESSION: THEORY

Theorem:

$$\mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_\lambda \|_2^2 \right] = \underbrace{\lambda \sum_{j=1}^p \frac{\lambda_j \lambda}{(\lambda_j + \lambda)^2} \langle \mathbf{v}_j, \mathbf{w}_\star \rangle^2}_{\text{bias}} + \underbrace{\frac{\sigma^2}{n} \sum_{j=1}^p \left(\frac{\lambda_j}{\lambda_j + \lambda} \right)^2}_{\text{variance}}.$$

Corollary: since $\lambda_j \lambda / (\lambda_j + \lambda)^2 \leq 1/2$ and $\sum_{j=1}^p \lambda_j = \text{tr}(\frac{1}{n} \mathbf{X}^\top \mathbf{X})$,

$$\mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_\lambda \|_2^2 \right] \leq \frac{\lambda \| \mathbf{w}_\star \|_2^2}{2} + \frac{\sigma^2 \text{tr}(\frac{1}{n} \mathbf{X}^\top \mathbf{X})}{2n\lambda}.$$

RIDGE REGRESSION: THEORY

Theorem:

$$\mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_\lambda \|_2^2 \right] = \underbrace{\lambda \sum_{j=1}^p \frac{\lambda_j \lambda}{(\lambda_j + \lambda)^2} \langle \mathbf{v}_j, \mathbf{w}_\star \rangle^2}_{\text{bias}} + \underbrace{\frac{\sigma^2}{n} \sum_{j=1}^p \left(\frac{\lambda_j}{\lambda_j + \lambda} \right)^2}_{\text{variance}}.$$

Corollary: since $\lambda_j \lambda / (\lambda_j + \lambda)^2 \leq 1/2$ and $\sum_{j=1}^p \lambda_j = \text{tr}(\frac{1}{n} \mathbf{X}^\top \mathbf{X})$,

$$\mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_\lambda \|_2^2 \right] \leq \frac{\lambda \| \mathbf{w}_\star \|_2^2}{2} + \frac{\sigma^2 \text{tr}(\frac{1}{n} \mathbf{X}^\top \mathbf{X})}{2n\lambda}.$$

For instance, using $\lambda := \sqrt{\frac{\text{tr}(\frac{1}{n} \mathbf{X}^\top \mathbf{X})}{n}}$ guarantees

$$\mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_\lambda \|_2^2 \right] \leq \frac{\| \mathbf{w}_\star \|_2^2 + \sigma^2}{2} \sqrt{\frac{\text{tr}(\frac{1}{n} \mathbf{X}^\top \mathbf{X})}{n}}.$$

RIDGE REGRESSION: THEORY

Theorem:

$$\mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_\lambda \|_2^2 \right] = \underbrace{\lambda \sum_{j=1}^p \frac{\lambda_j \lambda}{(\lambda_j + \lambda)^2} \langle \mathbf{v}_j, \mathbf{w}_\star \rangle^2}_{\text{bias}} + \underbrace{\frac{\sigma^2}{n} \sum_{j=1}^p \left(\frac{\lambda_j}{\lambda_j + \lambda} \right)^2}_{\text{variance}}.$$

Corollary: since $\lambda_j \lambda / (\lambda_j + \lambda)^2 \leq 1/2$ and $\sum_{j=1}^p \lambda_j = \text{tr}(\frac{1}{n} \mathbf{X}^\top \mathbf{X})$,

$$\mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_\lambda \|_2^2 \right] \leq \frac{\lambda \| \mathbf{w}_\star \|_2^2}{2} + \frac{\sigma^2 \text{tr}(\frac{1}{n} \mathbf{X}^\top \mathbf{X})}{2n\lambda}.$$

For instance, using $\lambda := \sqrt{\frac{\text{tr}(\frac{1}{n} \mathbf{X}^\top \mathbf{X})}{n}}$ guarantees

$$\mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_\lambda \|_2^2 \right] \leq \frac{\| \mathbf{w}_\star \|_2^2 + \sigma^2}{2} \sqrt{\frac{\text{tr}(\frac{1}{n} \mathbf{X}^\top \mathbf{X})}{n}}.$$

No explicit dependence on p . Corollary can be meaningful even if $p = \infty$, as long as $\| \mathbf{w}_\star \|_2^2$ and $\text{tr}(\frac{1}{n} \mathbf{X}^\top \mathbf{X})$ are finite.

PRINCIPAL COMPONENTS (“KEEP OR KILL”) REGRESSION

Ridge regression as shrinkage:

$$\hat{\mathbf{w}}_{\lambda} = \sum_{j=1}^p \frac{\lambda_j}{\lambda_j + \lambda} \langle \mathbf{v}_j, \hat{\mathbf{w}}_{\text{ols}} \rangle \mathbf{v}_j.$$

PRINCIPAL COMPONENTS (“KEEP OR KILL”) REGRESSION

Ridge regression as shrinkage:

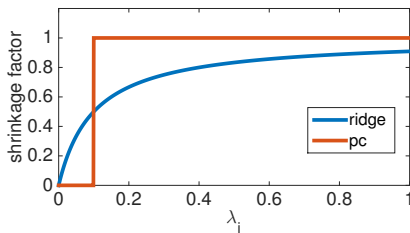
$$\hat{\mathbf{w}}_{\lambda} = \sum_{j=1}^p \frac{\lambda_j}{\lambda_j + \lambda} \langle \mathbf{v}_j, \hat{\mathbf{w}}_{\text{ols}} \rangle \mathbf{v}_j.$$

Another approach: principal components regression.

Instead of shrinking $\hat{\mathbf{w}}_{\text{ols}}$ in all directions,

- ▶ either keep $\hat{\mathbf{w}}_{\text{ols}}$ in direction $\mathbf{v}_j$ (if $\lambda_j \geq \lambda$),
- ▶ or kill $\hat{\mathbf{w}}_{\text{ols}}$ in direction $\mathbf{v}_j$ (if $\lambda_j < \lambda$).

$$\hat{\mathbf{w}}_{\text{pc } \lambda} := \sum_{j=1}^p \mathbb{1}\{\lambda_j \geq \lambda\} \langle \mathbf{v}_j, \hat{\mathbf{w}}_{\text{ols}} \rangle \mathbf{v}_j.$$



PRINCIPAL COMPONENTS REGRESSION: THEORY

Theorem:

$$\mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_{\text{pc } \lambda} \|_2^2 \right] = \underbrace{\sum_{j=1}^p \mathbb{1}\{\lambda_j < \lambda\} \lambda_j \langle \mathbf{v}_j, \mathbf{w}_\star \rangle^2}_{\text{bias}} + \underbrace{\frac{\sigma^2}{n} \sum_{j=1}^p \mathbb{1}\{\lambda_j \geq \lambda\}}_{\text{variance}}$$

PRINCIPAL COMPONENTS REGRESSION: THEORY

Theorem:

$$\begin{aligned}\mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_{\text{pc } \lambda} \|^2_2 \right] &= \underbrace{\sum_{j=1}^p \mathbb{1}\{\lambda_j < \lambda\} \lambda_j \langle \mathbf{v}_j, \mathbf{w}_\star \rangle^2}_{\text{bias}} + \underbrace{\frac{\sigma^2}{n} \sum_{j=1}^p \mathbb{1}\{\lambda_j \geq \lambda\}}_{\text{variance}} \\ &\leq 4 \cdot \mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_\lambda \|^2_2 \right].\end{aligned}$$

PRINCIPAL COMPONENTS REGRESSION: THEORY

Theorem:

$$\begin{aligned}\mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_{\text{pc } \lambda} \|_2^2 \right] &= \underbrace{\sum_{j=1}^p \mathbb{1}\{\lambda_j < \lambda\} \lambda_j \langle \mathbf{v}_j, \mathbf{w}_\star \rangle^2}_{\text{bias}} + \underbrace{\frac{\sigma^2}{n} \sum_{j=1}^p \mathbb{1}\{\lambda_j \geq \lambda\}}_{\text{variance}} \\ &\leq 4 \cdot \mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_\lambda \|_2^2 \right].\end{aligned}$$

- Should pick λ large enough so that

$$n \geq \sum_{j=1}^p \mathbb{1}\{\lambda_j \geq \lambda\}$$

(effective dimension at cut-off point λ).

PRINCIPAL COMPONENTS REGRESSION: THEORY

Theorem:

$$\begin{aligned}\mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_{\text{pc } \lambda} \|^2_2 \right] &= \underbrace{\sum_{j=1}^p \mathbb{1}\{\lambda_j < \lambda\} \lambda_j \langle \mathbf{v}_j, \mathbf{w}_\star \rangle^2}_{\text{bias}} + \underbrace{\frac{\sigma^2}{n} \sum_{j=1}^p \mathbb{1}\{\lambda_j \geq \lambda\}}_{\text{variance}} \\ &\leq 4 \cdot \mathbb{E} \left[\frac{1}{n} \| \mathbf{X} \mathbf{w}_\star - \mathbf{X} \hat{\mathbf{w}}_\lambda \|^2_2 \right].\end{aligned}$$

- Should pick λ large enough so that

$$n \geq \sum_{j=1}^p \mathbb{1}\{\lambda_j \geq \lambda\}$$

(effective dimension at cut-off point λ).

- Never (much) worse than ridge regression; **often substantially better**.

SPARSE REGRESSION AND LASSO

Another way to deal with $p > n$ is to only consider *sparse* $\boldsymbol{w}$ —i.e., $\boldsymbol{w}$ with only a small number ($\ll p$) of non-zero entries.

Another way to deal with $p > n$ is to only consider *sparse* w —i.e., w with only a small number ($\ll p$) of non-zero entries.

Other advantages of sparsity (especially relative to ridge/p.c.-regression):

- ▶ Sparse solutions more interpretable.
- ▶ Can be more efficient to evaluate $\langle x, w \rangle$ (both in terms of computing variable values and computing inner product).

For any $T \subseteq \{1, 2, \dots, p\}$, let $\hat{\mathbf{w}}(T) :=$ OLS only using variables in T .

Subset selection

Brute-force strategy. Pick the $T \subseteq \{1, 2, \dots, p\}$ of size $|T| = k$ for which

$$\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}(T)\|_2^2$$

is minimal, and return $\hat{\mathbf{w}}(T)$.

For any $T \subseteq \{1, 2, \dots, p\}$, let $\hat{\mathbf{w}}(T) :=$ OLS only using variables in T .

Subset selection

Brute-force strategy. Pick the $T \subseteq \{1, 2, \dots, p\}$ of size $|T| = k$ for which

$$\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}(T)\|_2^2$$

is minimal, and return $\hat{\mathbf{w}}(T)$.

Gives you exactly what you want (for given value k).

SPARSE REGRESSION METHODS

For any $T \subseteq \{1, 2, \dots, p\}$, let $\hat{\mathbf{w}}(T) :=$ OLS only using variables in T .

Subset selection

Brute-force strategy. Pick the $T \subseteq \{1, 2, \dots, p\}$ of size $|T| = k$ for which

$$\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}(T)\|_2^2$$

is minimal, and return $\hat{\mathbf{w}}(T)$.

Gives you exactly what you want (for given value k).

Only feasible for very small k , since complexity scales with $\binom{p}{k}$.
(NP-hard optimization problem.)

Forward stepwise regression

Greedy strategy. Starting with $T = \emptyset$, repeat until $|T| = k$:

Pick the $j \in \{1, 2, \dots, p\} \setminus T$ for which

$$\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}(T \cup \{j\})\|_2^2$$

is minimal, and add this j to T .

Return $\hat{\mathbf{w}}(T)$.

Forward stepwise regression

Greedy strategy. Starting with $T = \emptyset$, repeat until $|T| = k$:

Pick the $j \in \{1, 2, \dots, p\} \setminus T$ for which

$$\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}(T \cup \{j\})\|_2^2$$

is minimal, and add this j to T .

Return $\hat{\mathbf{w}}(T)$.

Gives you a k -sparse solution, no shrinkage within T .

SPARSE REGRESSION METHODS

Forward stepwise regression

Greedy strategy. Starting with $T = \emptyset$, repeat until $|T| = k$:

Pick the $j \in \{1, 2, \dots, p\} \setminus T$ for which

$$\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}(T \cup \{j\})\|_2^2$$

is minimal, and add this j to T .

Return $\hat{\mathbf{w}}(T)$.

Gives you a k -sparse solution, no shrinkage within T .

Primarily only effective when columns of $\mathbf{X}$ are close to orthogonal.

LASSO (TIBSHIRANI, 1994)

Lasso: least absolute shrinkage and selection operator

$$\hat{\mathbf{w}}_{\text{lasso } \lambda} := \arg \min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1.$$

(Convex, though not differentiable.)

LASSO (TIBSHIRANI, 1994)

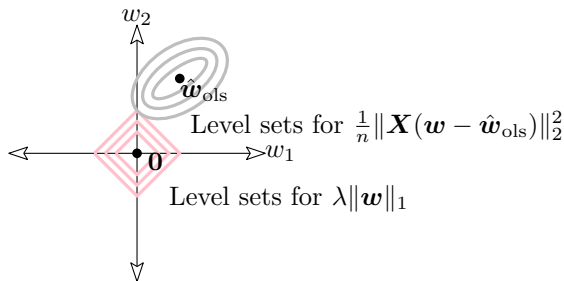
Lasso: least absolute shrinkage and selection operator

$$\hat{\mathbf{w}}_{\text{lasso } \lambda} := \arg \min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1.$$

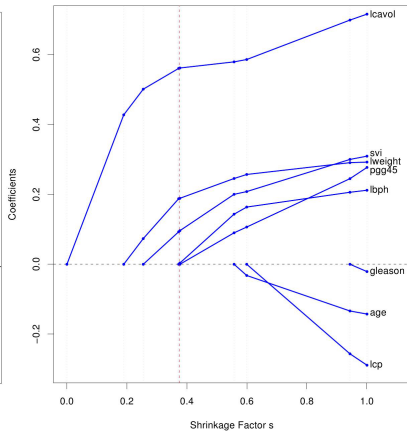
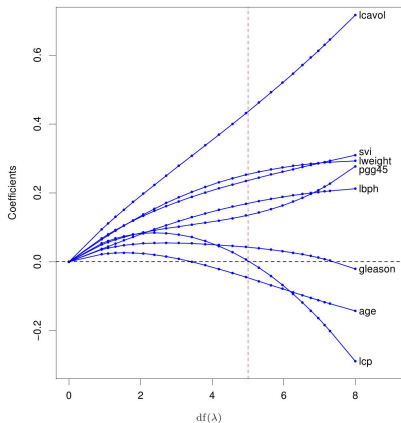
(Convex, though not differentiable.)

Objective (as function of $\mathbf{w}$) can be written as

$$\frac{1}{n} \|\mathbf{X}(\mathbf{w} - \hat{\mathbf{w}}_{\text{ols}})\|_2^2 + \lambda \|\mathbf{w}\|_1 + (\text{stuff not depending on } \mathbf{w})$$



COEFFICIENT PROFILE



Horizontal axis: varying λ (large λ to left, small λ to right).

Vertical axis: coefficient value in $\hat{w}_\lambda$ for eight different variables.

LASSO: THEORY

Many results, mostly roughly of the following flavor.

Suppose

- ▶ w_* has $\leq k$ non-zero entries;
- ▶ X satisfies some special properties (typically not efficiently checkable);
- ▶ $\lambda \approx \sigma \sqrt{\frac{2 \log(p)}{n}}$;

then

$$\mathbb{E} \left[\frac{1}{n} \|Xw_* - X\hat{w}_{\text{lasso } \lambda}\|_2^2 \right] \leq O\left(\frac{\sigma^2 k \log(p)}{n}\right).$$

LASSO: THEORY

Many results, mostly roughly of the following flavor.

Suppose

- ▶ w_* has $\leq k$ non-zero entries;
- ▶ X satisfies some special properties (typically not efficiently checkable);
- ▶ $\lambda \approx \sigma \sqrt{\frac{2 \log(p)}{n}}$;

then

$$\mathbb{E} \left[\frac{1}{n} \|Xw_* - X\hat{w}_{\text{lasso } \lambda}\|_2^2 \right] \leq O\left(\frac{\sigma^2 k \log(p)}{n}\right).$$

Very active subject of research; closely related to “compressed sensing”; intersects with beautiful subject of high-dimensional convex geometry.

- ▶ Fixed-design setting for studying linear regression methods.
- ▶ **Ridge and principal components regression** make use of eigenvectors of $\frac{1}{n} \mathbf{X}^\top \mathbf{X}$ (“principal component directions”), and also corresponding eigenvalues.
- ▶ **Ridge regression**: shrink $\hat{\mathbf{w}}_{\text{ols}}$ along principal component directions by amount related to eigenvalue and λ .
- ▶ **Principal components regression**: keep-or-kill $\hat{\mathbf{w}}_{\text{ols}}$ along principal component directions, based on comparing eigenvalue to λ .
- ▶ **Sparse regression**: intractable, but some greedy strategies work.
- ▶ **Lasso**: shrink coefficients towards zero in a way that tends to lead to sparse solutions.