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

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

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

Fixed-design linear regression

A simplified fixed-design setting

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

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

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

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

Fixed-design linear regression

A simplified fixed-design setting

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

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

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

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

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

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

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

Ordinary least squares

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

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

Ordinary least squares

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

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

Expressing $\boldsymbol{X}\hat{\boldsymbol{w}}_{\mathrm{ols}}$ in terms of $\boldsymbol{X}\boldsymbol{w}_\star$:

$$\boldsymbol{X}\hat{\boldsymbol{w}}_{\mathrm{ols}}$$

Ordinary least squares

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

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

Expressing $\boldsymbol{X}\hat{\boldsymbol{w}}_{\mathrm{ols}}$ in terms of $\boldsymbol{X}\boldsymbol{w}_\star$:

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

Ordinary least squares

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

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

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

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

Ordinary least squares

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

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

Expressing $\boldsymbol{X}\hat{\boldsymbol{w}}_{\mathrm{ols}}$ in terms of $\boldsymbol{X}\boldsymbol{w}_\star$:

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

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

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

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

Expressing $\boldsymbol{X}\hat{\boldsymbol{w}}_{\mathrm{ols}}$ in terms of $\boldsymbol{X}\boldsymbol{w}_\star$:

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

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

$$\mathbb{E}\left[\frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}_\star - \boldsymbol{X}\hat{\boldsymbol{w}}_{\mathrm{ols}}\|_2^2\right] = \mathbb{E}\left[\frac{1}{n}\|\boldsymbol{\Pi}\boldsymbol{\varepsilon}\|_2^2\right]$$

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

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

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

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

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

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}_\star - \boldsymbol{X}\hat{\boldsymbol{w}}_{\text{ols}}\|_2^2\right] &= \mathbb{E}\left[\frac{1}{n}\|\boldsymbol{\Pi}\boldsymbol{\varepsilon}\|_2^2\right] \\ &= \mathbb{E}\left[\frac{1}{n}\operatorname{tr}\big(\boldsymbol{\Pi}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big)\right] \end{aligned}$$

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

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

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

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

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

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}_\star - \boldsymbol{X}\hat{\boldsymbol{w}}_{\text{ols}}\|_2^2\right] &= \mathbb{E}\left[\frac{1}{n}\|\boldsymbol{\Pi}\boldsymbol{\varepsilon}\|_2^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\operatorname{tr}\big(\boldsymbol{\Pi}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big)\right] \\
&= \frac{1}{n}\operatorname{tr}\big(\boldsymbol{\Pi}\mathbb{E}\big(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big)\big)
\end{aligned}$$

Ordinary least squares

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

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

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

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

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

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}_\star - \boldsymbol{X}\hat{\boldsymbol{w}}_{\text{ols}}\|_2^2\right] &= \mathbb{E}\left[\frac{1}{n}\|\boldsymbol{\Pi}\boldsymbol{\varepsilon}\|_2^2\right] \\ &= \mathbb{E}\left[\frac{1}{n}\operatorname{tr}\big(\boldsymbol{\Pi}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big)\right] \\ &= \frac{1}{n}\operatorname{tr}\big(\boldsymbol{\Pi}\mathbb{E}\big(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big)\big) \\ &= \frac{\sigma^2}{n}\operatorname{tr}(\boldsymbol{\Pi}) \end{aligned}$$

Ordinary least squares

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

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

Expressing $\boldsymbol{X}\hat{\boldsymbol{w}}_{\mathrm{ols}}$ in terms of $\boldsymbol{X}\boldsymbol{w}_\star$:

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

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

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}_\star - \boldsymbol{X}\hat{\boldsymbol{w}}_{\mathrm{ols}}\|_2^2\right] &= \mathbb{E}\left[\frac{1}{n}\|\boldsymbol{\Pi}\boldsymbol{\varepsilon}\|_2^2\right] \\ &= \mathbb{E}\left[\frac{1}{n}\mathrm{tr}\big(\boldsymbol{\Pi}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big)\right] \\ &= \frac{1}{n}\mathrm{tr}\big(\boldsymbol{\Pi}\mathbb{E}\big(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big)\big) \\ &= \frac{\sigma^2}{n}\mathrm{tr}(\boldsymbol{\Pi}) = \frac{\sigma^2 p}{n}. \end{aligned}$$

note the p

???

Ridge regression and principal components regression

Ridge regression

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

Ridge regression

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

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

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

$$\hat{\boldsymbol{w}}_\lambda := \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2.$$

- Gradient of objective is zero when

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

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

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

Ridge regression

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

$$\hat{\boldsymbol{w}}_\lambda := \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^p} \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2.$$

- Gradient of objective is zero when

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

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

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

- $\hat{\boldsymbol{w}}_\lambda$ approaches $\hat{\boldsymbol{w}}_{\text{ols}}$ as $\lambda \to 0$.

when the row is
smaller than column

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

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

- Gradient of objective is zero when

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

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

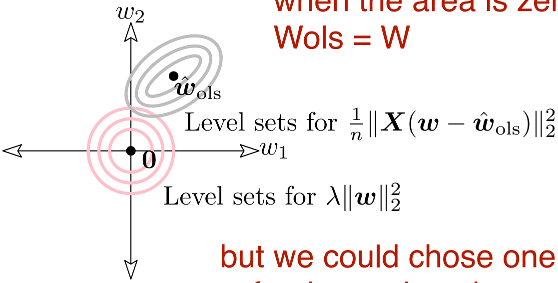
$$\hat{\boldsymbol{w}}_\lambda := \left(\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I} \right)^{-1} \left(\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{y} \right).$$

- $\hat{\boldsymbol{w}}_\lambda$ approaches $\hat{\boldsymbol{w}}_{\text{ols}}$ as $\lambda \to 0$.

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

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

when the area is zero is when Wols = W



but we could chose one w for the setting. how to make the trade off?

Ridge regression: eigendecomposition

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

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

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

Ridge regression: eigendecomposition

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

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

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

- Eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p$ comprise an *orthonormal basis* for $\mathbb{R}^p$. We'll look at $\hat{\boldsymbol{w}}_{\text{ols}}$ and $\boldsymbol{w}_\star$ in this basis: e.g.,

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

Ridge regression: eigendecomposition

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

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

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

- Eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p$ comprise an *orthonormal basis* for $\mathbb{R}^p$. We'll look at $\hat{\boldsymbol{w}}_{\text{ols}}$ and $\boldsymbol{w}_\star$ in this basis: e.g.,

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

the value along vj direction

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

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

Ridge regression vs. ordinary least squares

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

If $\hat{\boldsymbol{w}}_{\mathrm{ols}}$ exists, then

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

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

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

If $\hat{\boldsymbol{w}}_{\mathrm{ols}}$ exists, then

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

If $\hat{\boldsymbol{w}}_{\mathrm{ols}}$ exists, then

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

Ridge regression vs. ordinary least squares

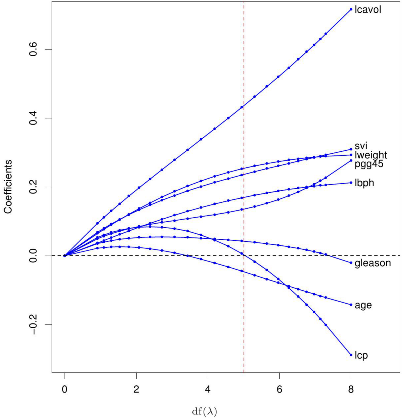
If $\hat{\boldsymbol{w}}_{\mathrm{ols}}$ exists, then

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

change the sharp
of Wols ?

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

Coefficient profile



note: the coefficient could could be used to decide direction. direction should be decide by vector.

Horizontal axis: varying λ (large λ to left, small λ to right).
Vertical axis: coefficient value in $\hat{\boldsymbol{w}}_\lambda$ for eight different variables.

Theorem:

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

Ridge regression: theory

Theorem:

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

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

$$\mathbb{E}\left[\frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}_\star - \boldsymbol{X}\hat{\boldsymbol{w}}_\lambda\|_2^2\right] \;\le\; \frac{\lambda\|\boldsymbol{w}_\star\|_2^2}{2} + \frac{\sigma^2\,\mathrm{tr}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)}{2n\lambda}.$$

Ridge regression: theory

Theorem:

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

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

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

For instance, using $\lambda := \sqrt{\frac{\operatorname{tr}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)}{n}}$ guarantees

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

Ridge regression: theory

Theorem:

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

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

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

this is super import! exclude the effects from p!!! (p >n)

For instance, using $\lambda := \sqrt{\frac{\mathrm{tr}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)}{n}}$ guarantees

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

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

Ridge regression as shrinkage:

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

Ridge regression as shrinkage:

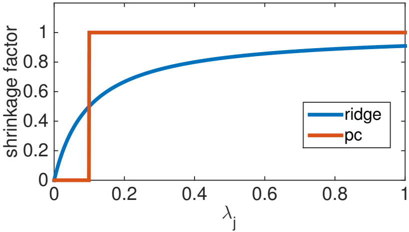
$$\hat{\boldsymbol{w}}_\lambda = \boxed{\sum_{j=1}^{p} \frac{\lambda_j}{\lambda_j + \lambda} \langle \boldsymbol{v}_j, \hat{\boldsymbol{w}}_{\text{ols}} \rangle \boldsymbol{v}_j .}$$

only
\lambda j

Another approach: principal components regression.
Instead of shrinking $\hat{\boldsymbol{w}}_{\text{ols}}$ in all directions,

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

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



Theorem:

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

Principal components regression: theory

Theorem:

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

Theorem:

$$\mathbb{E}\left[\frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}_\star - \boldsymbol{X}\hat{\boldsymbol{w}}_{\mathrm{pc}\,\lambda}\|_2^2\right] = \sum_{j=1}^{p} \mathbb{1}\{\lambda_j < \lambda\}\lambda_j\langle \boldsymbol{v}_j, \boldsymbol{w}_\star\rangle^2 + \frac{\sigma^2}{n}\sum_{j=1}^{p}\mathbb{1}\{\lambda_j \geq \lambda\}$$
$$\leq 4 \cdot \mathbb{E}\left[\frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}_\star - \boldsymbol{X}\hat{\boldsymbol{w}}_\lambda\|_2^2\right].$$

- 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:

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

- Should pick λ large enough so that

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

  (effective dimension at cut-off point λ).
- Never (much) worse than ridge regression; often substantially better.

Sparse regression and Lasso

Sparsity

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.

Sparsity

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.

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

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

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

Subset selection

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

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

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

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

Subset selection

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

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

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

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

Sparse regression methods

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

Subset selection

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

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

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

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 $j \in \{1, 2, \ldots, p\} \setminus T$ for which

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

is minimal, and add this j to T.

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

Forward stepwise regression

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

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

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

is minimal, and add this j to T.

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

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

Forward stepwise regression

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

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

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

is minimal, and add this j to T.

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

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

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

Lasso (Tibshirani, 1994)

Lasso: least absolute shrinkage and selection operator

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

(Convex, though not differentiable.)

Lasso (Tibshirani, 1994)

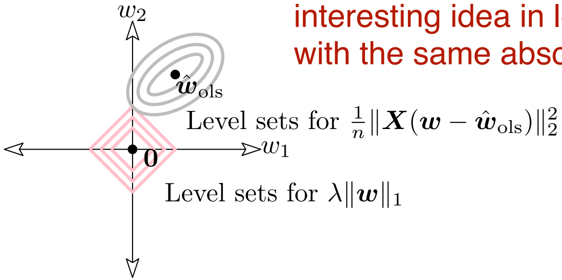
Lasso: least absolute shrinkage and selection operator

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

(Convex, though not differentiable.)

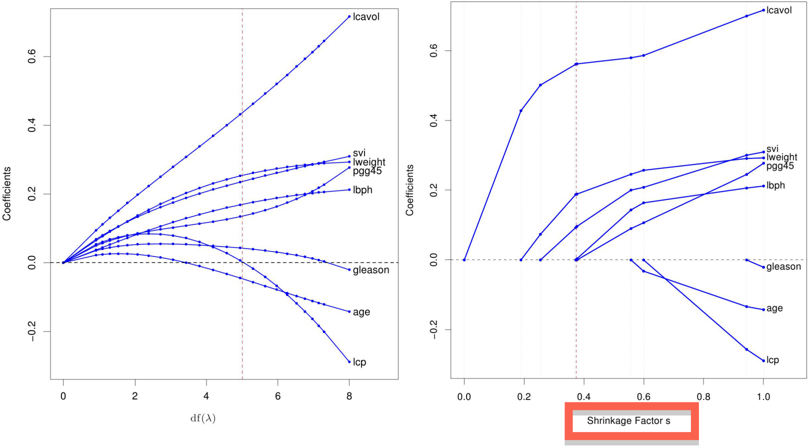
Objective (as function of $\boldsymbol{w}$) can be written as

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



interesting idea in level sets!!!
with the same absolute value

for different variables, for the same change in r



Horizontal axis: varying λ (large λ to left, small λ to right).
Vertical axis: coefficient value in $\hat{\boldsymbol{w}}_\lambda$ for eight different variables.

note: each has its own w

Lasso: theory

Many results, mostly roughly of the following flavor.
Suppose

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

then

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

Lasso: theory

Many results, mostly roughly of the following flavor.
Suppose

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

then

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

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

Recap

- Fixed-design setting for studying linear regression methods.
- **Ridge and principal components regression** make use of eigenvectors of $\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}$ ("principal component directions"), and also corresponding eigenvalues.
- **Ridge regression**: shrink $\hat{\boldsymbol{w}}_{\mathrm{ols}}$ along principal component directions by amount related to eigenvalue and λ.
- **Principal components regression**: keep-or-kill $\hat{\boldsymbol{w}}_{\mathrm{ols}}$ 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.