MVE137
Probability and Statistical Learning Using Python

Linear methods for regression

Alexandre Graell i Amat

alexandre.graell@chalmers.se

https://sites.google.com/site/agraellamat

September 23 and 27, 2022

CHALMERS

Linear regression

Linear regression: Assumes linear model for $f(\boldsymbol{x})$,

$$\hat{y} = f(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{w}\,,$$

i.e., $\tilde{f}(\boldsymbol{x}) = \mathbb{E}_{\mathsf{y}|\mathbf{x}}[y|x] \approx \tilde{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{w}^*$

- Simple and interpretable
- Can outperform non-linear methods when small number of training samples, very noisy data, sparse data
- Can be used to model nonlinear relationships using basis functions $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_M(\boldsymbol{x}))$
- Many nonlinear techniques direct generalizations of linear methods

Linear regression IBM

Linear regression and least squares

We assume:

- A probabilistic model

$$y = \tilde{f}(\boldsymbol{x}) + \varepsilon$$

  with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
- A linear model

$$f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T} \boldsymbol{w}\,,$$

  with $\boldsymbol{w} = (w_0, w_1, \ldots, w_p)^\mathsf{T}$ and $\boldsymbol{x} = (1, x_1, \ldots, x_p)^\mathsf{T}$
- A data set $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\} = (\boldsymbol{X}_{N\times(d+1)}, \boldsymbol{y}_{N\times 1})$
- Residual sum of squares criterion:

$$\mathrm{RSS}(\boldsymbol{w}) = \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}\right)^2$$

Linear regression and least squares

Optimal solution:

$$\boldsymbol{w}^* = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$$

obtained from differentiating the RSS and setting derivative to zero,

$$\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = 0$$

Predicted value corresponding to $\boldsymbol{x} = (1, x_1, \ldots, x_d)$:

$$\hat{y} = \boldsymbol{x}^\mathsf{T}\boldsymbol{w}^*$$

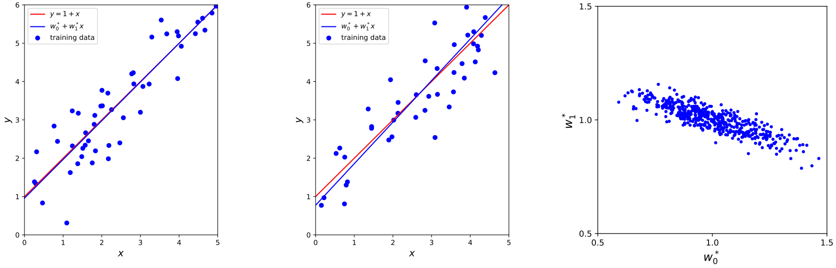
Properties of $\boldsymbol{w}^*$

$\boldsymbol{w}^*$ depends on the training data $\longrightarrow$ $\boldsymbol{w}^*$ is a random variable!

Example:

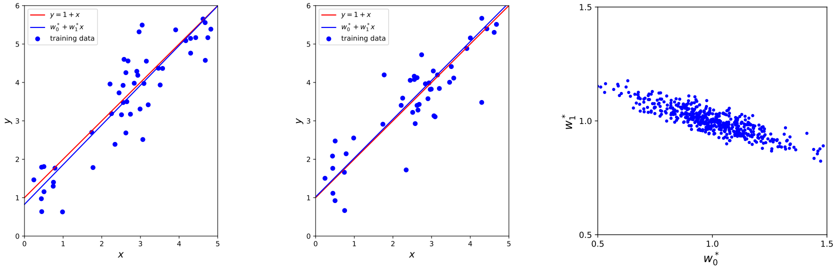
$$y = w_0 + w_1 x + \varepsilon\,,$$

with $\boldsymbol{w} = (1,1)^\mathsf{T}$, i.e., $y = 1 + x + \varepsilon$, and $\varepsilon = \mathcal{N}(0, 0.49)$



Changing training input data

Properties of $\boldsymbol{w}^*$



Fixed training input data

Distribution of $\boldsymbol{w}^*$

Assumptions:

- $y = \boldsymbol{x}^\mathsf{T}\boldsymbol{w} + \varepsilon$ (for training set: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\varepsilon}$)
- Observations y_i uncorrelated and with constant variance σ^2
- $\boldsymbol{X}$ fixed

Expectation:

$$\begin{aligned}\mathbb{E}[\boldsymbol{w}^*] &= \mathbb{E}\left[(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\mathbf{y}\right] \\ &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\mathbb{E}\left[\mathbf{y}\right]\end{aligned}$$

Fixed $\boldsymbol{X}$: $\mathbb{E}[\mathbf{y}] = \boldsymbol{X}\boldsymbol{w}$

Thus,

$$\begin{aligned}\mathbb{E}[\boldsymbol{w}^*] &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{w} \\ &= \boldsymbol{w}\end{aligned}$$

Distribution of $\boldsymbol{w}^*$

Variance:

$$\begin{aligned}\boldsymbol{w}^* - \mathbb{E}[\boldsymbol{w}^*] &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} - \boldsymbol{w} \\ &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} - (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})\boldsymbol{w} \\ &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\ &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{\varepsilon}\end{aligned}$$

Then:

$$\begin{aligned}\mathsf{Var}[\boldsymbol{w}^*] &= \mathbb{E}\left[(\boldsymbol{w}^* - \mathbb{E}[\boldsymbol{w}^*])(\boldsymbol{w}^* - \mathbb{E}[\boldsymbol{w}^*])^\mathsf{T}\right] \\ &= \mathbb{E}\left[(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\mathsf{T}\boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\right] \\ &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\mathsf{Var}[\boldsymbol{\varepsilon}]\boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}(\sigma^2\boldsymbol{I}_N)\boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\sigma^2\end{aligned}$$

Distribution of $\boldsymbol{w}^*$

And σ^2? Estimate it by the sample variance!

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

We typically use

$$\hat{\sigma}^2 = \frac{1}{N-d-1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

so that $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.

Distribution of $\boldsymbol{w}^*$

Distribution:

From $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\varepsilon}$ and $\boldsymbol{w}^* = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$,

$$\begin{aligned}\boldsymbol{w}^* &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}(\boldsymbol{X}\boldsymbol{w} + \boldsymbol{\varepsilon})\\ &= \boldsymbol{w} + (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{\varepsilon}\end{aligned}$$

$\boldsymbol{w}^*$ is a linear transformation of a multivariate Gaussian ($\boldsymbol{\varepsilon}$) $\longrightarrow$ $\boldsymbol{w}^*$ multivariate Gaussian!

$$\boldsymbol{w}^* \sim \mathcal{N}(\boldsymbol{w}, (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\sigma^2)$$

Interpretability of the model

Machine learning algorithms often a black box

Explainable AI: Understand decisions or predictions made by the AI

Linear regression allows for interpretability!

Interpretation of the estimated weights

$$\boldsymbol{w}^* \sim \mathcal{N}(\boldsymbol{w}, (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\sigma^2)$$

Interpretation of $(w_0^*, \ldots, w_d^*)$ estimated by least squares?

- Which w's are probably zero? $\longrightarrow$ associated features are irrelevant

If true $w_j = 0$:

$$w_j^* \sim \mathcal{N}(0, \sigma^2 v_j)$$

- If $w_j^* > \sigma^2 v_j$: highly improbable $w_j = 0$

Test hypothesis $w_j = 0$ (z-score):

$$z_j = \frac{w_j^*}{\hat{\sigma}\sqrt{v_j}}$$

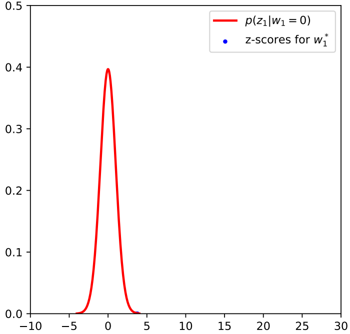
- $w_j = 0$: z_j has a student's t-distribution with $N - d - 1$ degrees of freedom

Interpretation of the estimated weights

Example:

$$y = w_0 + w_1 x + \varepsilon\,,$$

with $\boldsymbol{w} = (1, 1)^\mathsf{T}$, i.e., $y = 1 + x + \varepsilon$, and $\varepsilon = \mathcal{N}(0, 0.49)$



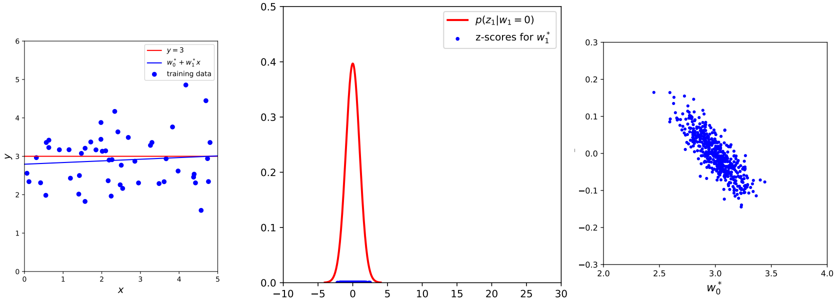
- $N = 50$ training examples
- blue circles: z-scores for w_1^* for 500 training data sets

Interpretation of the estimated weights

Example:

$$y = w_0 + w_1 x + \varepsilon\,,$$

with $\boldsymbol{w} = (3,0)^\mathsf{T}$, i.e., $y = 1 + x + \varepsilon$, and $\varepsilon = \mathcal{N}(0, 0.49)$



- $N = 50$ training examples
- $N = 500$ different training sets (right)

The Gauss-Markov theorem

The least squares estimate $\boldsymbol{w}^*$ has the smallest variance among all linear unbiased estimates.

We consider estimation of $\theta = \boldsymbol{a}^\mathsf{T}\boldsymbol{w}$. Least squares estimate:

$$\theta^* = \boldsymbol{a}^\mathsf{T}\boldsymbol{w}^* = \boldsymbol{a}^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$$

Assuming model $y = \boldsymbol{x}^\mathsf{T}\boldsymbol{w} + \varepsilon$ is correct $\longrightarrow \mathbb{E}[\mathbf{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{w}$ and

$$\begin{aligned}\mathbb{E}[\boldsymbol{a}^\mathsf{T}\boldsymbol{w}^*] &= \mathbb{E}\left[\boldsymbol{a}^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\mathbf{y}\right] \\ &= \boldsymbol{a}^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\mathbb{E}[\mathbf{y}] \\ &= \boldsymbol{a}^\mathsf{T}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{w} \\ &= \boldsymbol{a}^\mathsf{T}\boldsymbol{w} \\ &= \theta\end{aligned}$$

Gauss-Markov theorem. Any other linear estimator $\tilde{\theta} = \boldsymbol{c}^\mathsf{T}\boldsymbol{y}$ that is unbiased for $\boldsymbol{a}^\mathsf{T}\boldsymbol{w}$ has variance

$$\mathsf{Var}[\boldsymbol{a}^\mathsf{T}\boldsymbol{w}^*] \leq \mathsf{Var}[\boldsymbol{c}^\mathsf{T}\boldsymbol{y}]$$

The Gauss-Markov theorem: Implications

Mean-squared error of estimator $\tilde{\theta}$ of θ,

$$\begin{aligned}
\mathsf{MSE}(\tilde{\theta}) &= \mathbb{E}\left[(\tilde{\theta}-\theta)^2\right] \\
&= \mathbb{E}[(\tilde{\theta}-\mathbb{E}[\tilde{\theta}]+\mathbb{E}[\tilde{\theta}]-\theta)^2] \\
&= (\tilde{\theta}-\mathbb{E}[\tilde{\theta}])^2 + 2(\tilde{\theta}-\mathbb{E}[\tilde{\theta}])(\mathbb{E}[\tilde{\theta}]-\theta) + (\mathbb{E}[\tilde{\theta}]-\theta)^2 \\
&= (\tilde{\theta}-\mathbb{E}[\tilde{\theta}])^2 + (\mathbb{E}[\tilde{\theta}]-\theta)^2 \\
&= \underbrace{\mathsf{Var}[\tilde{\theta}]}_{\text{variance}} + (\underbrace{\mathbb{E}[\tilde{\theta}]-\theta}_{\text{bias}})^2
\end{aligned}$$

The Gauss-Markov theorem: Implications

Mean-squared error of estimator $\tilde{\theta}$ of θ,

$$\mathrm{MSE}(\tilde{\theta}) = \underbrace{\mathrm{Var}[\tilde{\theta}]}_{\text{variance}} + (\underbrace{\mathbb{E}[\tilde{\theta}] - \theta}_{\text{bias}})^2$$

Gauss-Markov theorem. For all linear estimators with zero bias, the least-squares estimator has the smallest MSE!

Biased estimates may give a smaller MSE $\longrightarrow$ bias-variance trade-off!

In practice: Any model will be biased $\longrightarrow$ strike the right bias-variance trade-off.

MSE and prediction error

Prediction error:

$$L(\hat{y}) = \sigma^2 + (\underbrace{\tilde{f}(\boldsymbol{x}) - \mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x})]}_{\text{bias}})^2 + \mathsf{Var}_{\mathcal{D}}\left[f(\boldsymbol{x})\right]$$

With $f(\boldsymbol{x}) \equiv \tilde{\theta}$ and $\tilde{f}(\boldsymbol{x}) \equiv \theta$:

$$\begin{aligned} L(\hat{y}) &= \sigma^2 + (\underbrace{\theta - \mathbb{E}[\tilde{\theta}]}_{\text{bias}})^2 + \mathsf{Var}_{\mathcal{D}}\left[\tilde{\theta}\right] \\ &= \sigma^2 + \mathsf{MSE}(\tilde{\theta}) \\ &= \sigma^2 + \mathsf{MSE}(f(\boldsymbol{x})) \end{aligned}$$

- σ^2: independent of model, irreducible
- $\mathsf{MSE}(f(\boldsymbol{x}))$: error in the model

Minimizing the MSE minimizes the expected prediction error

Multiple outputs

Predict $K > 1$ output variables $\boldsymbol{y} = (y_1, \ldots, y_K)^\mathsf{T}$

Linear model for each output:

$$
\begin{aligned}
y_j &= w_{j,0} + \sum_{\ell=1}^{d} x_\ell w_{j,\ell} + \varepsilon_i \\
&= w_{j,0} + \boldsymbol{w}_j^\mathsf{T} \boldsymbol{x} + \varepsilon_i \\
&= f_j(\boldsymbol{x}) + \varepsilon_i
\end{aligned}
$$

- $\mathcal{D} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$
- $\boldsymbol{Y}$: $N \times K$ matrix of outputs $\boldsymbol{Y}$, with $\boldsymbol{y}_i$ as i-th row

Linear regression:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W} + \boldsymbol{E}$$

$\boldsymbol{W}$: $(d+1) \times K$ matrix of coefficients w

Multiple outputs

Single output:

$$\mathsf{RSS}(\boldsymbol{w}) = \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{w}\right)^2$$

Multiple outputs:

$$\begin{aligned}
\mathsf{RSS}(\boldsymbol{W}) &= \sum_{i=1}^{N} \sum_{j=1}^{K} (y_{i,j} - f_j(\boldsymbol{x}_i))^2 \\
&= \sum_{i=1}^{N} \sum_{j=1}^{K} \left(y_{i,j} - \left(w_{j,0} + \sum_{\ell=1}^{d} x_{i,\ell} w_{j,\ell}\right)\right)^2 \\
&= \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_{\mathsf{F}}^2
\end{aligned}$$

with

$$\|\boldsymbol{A}\|_{\mathsf{F}}^2 = \sum_i \sum_j a_{i,j}^2$$

Multiple outputs

Least squares solution:

$$\begin{aligned}\boldsymbol{W}^* &= \arg\min_{\boldsymbol{W}} \ \mathsf{RSS}(\boldsymbol{W}) \\ &= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{Y}\end{aligned}$$

and

$$\hat{\boldsymbol{Y}}^* = \boldsymbol{X}\boldsymbol{W}^* = \boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{Y}$$

Coefficients for i-th output $y_{1,i},\ldots,y_{N,i}$ solution to least squares regression of $y_{1,i},\ldots,y_{N,i}$ on columns of $\boldsymbol{X}$ $\longrightarrow$ multiple outputs do not affect one another's estimates.

Subset selection

Least squares estimates:

- Low bias
- High variance: may penalize prediction accuracy

Accuracy can be improved by setting some w_j to zero.

Subset selection: Determine a subset of $k < d$ features which is the most informative $\longrightarrow$ Increased interpretability, improved accuracy.

Best-subset selection

For each $k \in \{1, \ldots, d\}$ find subset of size k that gives smallest RSS:

$$\mathsf{RSS}_{\mathsf{BSS}}(j_1, \ldots, j_k) = \min_{w_0, w_{j_1} \ldots, w_{j_k}} \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{\ell=1}^{k} w_{j_\ell} x_{i,j_\ell} \right)^2$$

with $j_i \in \{1, \ldots, d\}$

Observations:

- $\binom{d}{k}$ different subsets
- $d \leq 40$: computationally feasible algorithms
- $d > 40$: infeasible to solve exactly, heuristic algorithms
- How to choose k? bias-variance trade-off

Forward- and backward-stepwise selection

Idea: consider a greedy approach

Forward-stepwise selection:

Start with intercept, then augment model sequentially by adding predictor that improves fit most:

$$j_\ell = \arg\min_{j\in\mathcal{I}} \min_{w_0, w_{j_1}, \ldots, w_{j_{\ell-1}}, w_j} \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{m=1}^{\ell-1} w_{j_m} x_{i,j_m} - w_j x_{i,j} \right)^2$$

with $\mathcal{I} = \{1, \ldots, d\} \backslash \{j_1, \ldots, j_{\ell-1}\}$.

Backward-stepwise selection:

Start with complete model, and sequentially delete predictor that contributes least to fit $\longrightarrow$ At each step, we remove variable with smallest z-score

Subset selection

Principle: Set some w_j's to 0, let others be estimated using least squares.

- Interpretable
- Probably lower prediction

Drawbacks:

- Discrete procedure: w_j retained or discarded $\longrightarrow$ high variance
- High complexity (and heuristic solutions suboptimal)

Idea (Regularization): Allow all estimates to be positive, but constrain them to not become too big.

Shrinkage methods: Ridge regression

Idea: Shrink regression coefficients w_j by imposing a penalty on their size.

$$\boldsymbol{w}^*_{\text{ridge}} = \arg\min_{\boldsymbol{w}} \underbrace{\sum_{i=1}^{N}\left(y_i - w_0 - \sum_{i=1}^{d} x_{i,j} w_j\right)^2}_{\text{RSS}} + \lambda \underbrace{\sum_{j=1}^{d} w_j^2}_{\text{reg. function}}$$

$$= \arg\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda \|\tilde{\boldsymbol{w}}\|^2$$

Observations:

- Larger λ: more shrinkage
- $\lambda = \infty$: $\tilde{\boldsymbol{w}}^*_{\text{ridge}} = 0$
- $\lambda = 0$: conventional linear regression

Shrinkage methods: Ridge regression

Ridge regression in an equivalent form:

$$\boldsymbol{w}^*_{\text{ridge}} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{i=1}^{d} x_{i,j} w_j \right)^2$$
$$\text{subject to } \|\tilde{\boldsymbol{w}}\|^2 \leq t$$

Shrinkage methods: Ridge regression

Ridge regression in another equivalent form:

$$\boldsymbol{w}_{\mathsf{c}}^* = \arg\min_{\boldsymbol{w}_{\mathsf{c}}} \sum_{i=1}^{N} \left(y_i - w_{\mathsf{c},0} - \sum_{i=1}^{d} \tilde{x}_{i,j} w_{\mathsf{c},j} \right)^2 + \lambda \sum_{j=1}^{d} w_{\mathsf{c},j}^2 \,,$$

with

$$\tilde{x}_{i,j} = x_{i,j} - \bar{x}_j = x_{i,j} - \frac{1}{N} \sum_{\ell=1}^{N} x_{\ell,j}$$

Optimization can be done in two steps

1. Fit $w_{\mathsf{c},0}$ separately,

$$w_{\mathsf{c},0}^* = \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

2. Optimize all other $w_{\mathsf{c},j}$ by ridge regression without intercept

Shrinkage methods: Ridge regression

Assuming centered inputs $\tilde{x}_{i,j} = x_{i,j} - \bar{x}_j$ and $\tilde{y}_i = y_i - \bar{y}$, $w_{\mathsf{c},j}$, $j = 1, \ldots, d$, obtained solving

$$\begin{aligned}
\boldsymbol{w}_{\mathsf{c}}^* &= \arg\min_{\boldsymbol{w}_{\mathsf{c}}} \sum_{i=1}^{N} \left(y_i - \sum_{i=1}^{d} x_{i,j} w_{\mathsf{c},j} \right)^2 + \lambda \sum_{j=1}^{d} w_{\mathsf{c},j}^2 \\
&= \arg\min_{\boldsymbol{w}_{\mathsf{c}}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_{\mathsf{c}})^{\mathsf{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_{\mathsf{c}}) + \lambda \|\boldsymbol{w}_{\mathsf{c}}\|^2
\end{aligned}$$

with $\boldsymbol{y} = (y_1, \ldots, y_N)^{\mathsf{T}}$, $\boldsymbol{w}_{\mathsf{c}} = (w_{\mathsf{c},1}, \ldots, w_{\mathsf{c},d})^{\mathsf{T}}$, and

$$\boldsymbol{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,d} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,d} \\ \vdots & \vdots & \ldots & \vdots \\ x_{N,1} & x_{N,2} & \ldots & x_{N,d} \end{pmatrix}$$

$$\boldsymbol{w}_{\text{ridge}}^* = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^{\mathsf{T}} \boldsymbol{y}$$

Shrinkage methods: Ridge regression

We assume centered input matrix $\boldsymbol{X}$ (no intercept)

Singular value decomposition of $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}}$$

$\boldsymbol{U}$: $N \times d$ semi-orthogonal matrix; columns span column space of $\boldsymbol{X}$
$\boldsymbol{V}$: $d \times d$ orthogonal matrix
$\boldsymbol{D}$: $d \times d$ diagonal matrix with diagonal entries $d_1 \geq d_2 \geq \ldots \geq d_d$ singular values of $\boldsymbol{X}$

Can write least squares fitted vector as:

$$\begin{aligned}
\boldsymbol{X}\boldsymbol{w}^{\mathsf{ls}} &= \boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} \\
&= \boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y} \\
&= \sum_{i=1}^{d} \boldsymbol{u}_i(\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{y})
\end{aligned}$$

Shrinkage methods: Ridge regression

Least squares solution:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w}^{\mathsf{ls}} = \boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y}$$
$$= \sum_{i=1}^{d} \boldsymbol{u}_i(\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{y})$$

Observation:

- $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y}$: coordinates of $\boldsymbol{y}$ with respect to $\boldsymbol{U}$ $\longrightarrow$ least squares solution $\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y}$ is closest approximation to $\boldsymbol{y}$ in subspace spanned by columns of $\boldsymbol{U}$

Shrinkage methods: Ridge regression

Least squares solution:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w}^{\mathsf{ls}} = \boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y} = \sum_{i=1}^{d} \boldsymbol{u}_i(\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{y})$$

Ridge regression solution:

$$\begin{aligned}\hat{\boldsymbol{y}}_{\mathsf{ridge}} &= \boldsymbol{X}\boldsymbol{w}^*_{\mathsf{ridge}} \\ &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y} \\ &= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y}\end{aligned}$$

Observations:

- $\boldsymbol{D}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}$: diagonal matrix with elements $\frac{d_i^2}{d_i^2+\lambda}$
- $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y}$: coordinates of vector $\boldsymbol{y}$ in the basis spanned by $\boldsymbol{U}$

$$\hat{\boldsymbol{y}}_{\mathsf{ridge}} = \sum_{i=1}^{d} \boldsymbol{u}_i \frac{d_i^2}{d_i^2 + \lambda} \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{y}$$

Principal components and meaning of the d_i^2's

- A collection of points in a $\mathbb{R}^d$ (d-dimensional vectors)
- Want to summarize them by projecting them onto a q-dimensional subspace

Most informative summary: directions $\boldsymbol{b} \in \mathbb{R}^d$ with large spread of the data

Principal components: q orthogonal directions in space along which projections of the original vectors have largest variance.

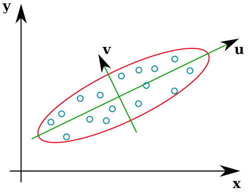
Principal components: q orthogonal directions that minimize average (mean-squared) distance between original vectors and their projections.

Principal components and meaning of the d_i^2's

Principal components: Sequence of q (unit) vectors, with i-th vector the direction of a line that best fits data while being orthogonal to first $i-1$ vectors

Best-fitting line $\equiv$ one that minimizes average squared distance from points to line $\equiv$ maximizes variance

1. 1st PC: direction in space along which projections have largest variance
2. 2nd PC: direction maximizing variance among all directions orthogonal to 1st PC
3. k-th PC: variance-maximizing direction orthogonal to previous $k-1$ components



Principal components and meaning of the d_i^2's

How do we find the principal components?

- Direction with largest variation: Eigenvector with largest eigenvalue
- Second direction with largest variation: Eigenvector with second largest eigenvalue
- ...

Back to our regression problem (with centered $\boldsymbol{X}$):

Sample covariance matrix:

$$\boldsymbol{\Sigma} = \frac{1}{N}\boldsymbol{X}^\mathsf{T}\boldsymbol{X} = \frac{1}{N}\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\mathsf{T}$$

The eigenvectors $\boldsymbol{v}_i$ (columns of $\boldsymbol{V}$) are the principal components of $\boldsymbol{X}$!

Diagonal entries of $\boldsymbol{D}^2$, $d_1^2 \geq d_2^2 \geq \ldots \geq d_d^2$, are the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^\mathsf{T}$

Principal components and meaning of the d_i^2's

Project $\boldsymbol{x}_i$'s onto first principal component $(\boldsymbol{v}_1)$:

$$z_i^{(1)} = \boldsymbol{v}_1^\mathsf{T}\boldsymbol{x}_i$$

Variance of $z_i^{(1)}$:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\left(z_i^{(1)}\right)^2 &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{v}_1^\mathsf{T}\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{v}_1 \\
&= \frac{1}{N}\boldsymbol{v}_1^\mathsf{T}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{v}_1 \\
&= \frac{1}{N}\boldsymbol{v}_1^\mathsf{T}\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\mathsf{T}\boldsymbol{v}_1 \\
&= \frac{d_1^2}{N}
\end{aligned}$$

We have: variance $z_i^{(1)} >$ variance $z_i^{(2)}$, ... variance $z_i^{(j)} >$ variance $z_i^{(j+1)}$

Small d_j's correspond to directions having small variance. Ridge regression shrinks these directions the most!

Shrinkage methods: The lasso

Idea: Shrink regression coefficients w_j by imposing a penalty on their size.

$$\boldsymbol{w}^*_{\text{lasso}} = \arg\min_{\boldsymbol{w}} \underbrace{\sum_{i=1}^{N} \left(y_i - w_0 - \sum_{i=1}^{d} x_{i,j} w_j \right)^2}_{\text{RSS}} + \lambda \underbrace{\sum_{j=1}^{d} |w_j|}_{\text{reg. function}}$$

$$= \arg\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda \|\tilde{\boldsymbol{w}}\|_1$$

Equivalently,

$$\boldsymbol{w}^*_{\text{lasso}} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{i=1}^{d} x_{i,j} w_j \right)^2$$

$$\text{subject to } \|\tilde{\boldsymbol{w}}\|_1 \leq t$$

Shrinkage methods: The lasso

$$\boldsymbol{w}^*_{\text{lasso}} = \arg\min_{\boldsymbol{w}} \underbrace{\sum_{i=1}^{N}\left(y_i - w_0 - \sum_{i=1}^{d} x_{i,j} w_j\right)^2}_{\text{RSS}} + \lambda \underbrace{\sum_{j=1}^{d} |w_j|}_{\text{reg. function}}$$

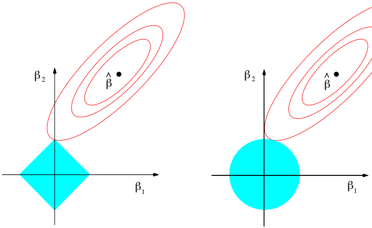
$$= \arg\min_{\boldsymbol{w}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda \|\tilde{\boldsymbol{w}}\|_1$$

Observations:

- Solution not a simple linear function of the y_i's
- No closed-form solution
- Computing the Lasso solution a quadratic programming problem
- We can standardize predictors: intercept can be fitted separately, then other parameters via lasso without intercept

Property: For λ sufficiently large, some coefficients w_j driven to zero $\longrightarrow$ can be interpreted as continuous subset selection

Ridge regression and the lasso: Graphical representation



- Elliptical contours: pairs (w_1, w_2) that yield a given RSS
- Ridge (right): $\|\boldsymbol{w}\|^2 \leq t \longrightarrow w_1^2 + w_1^2 \leq t$
- Lasso (left): $\|\boldsymbol{w}\|_1 \leq t \longrightarrow |w_1| + |w_2| \leq t$

Solution lies at intersection between RSS contours and constrained function.

- $\lambda = 0$: least-squares solution
- $\lambda = \infty$: $(0, 0)$
- Lasso: Intersection may occur in a vertex $\longrightarrow$ many zero coefficients

Generalization: Regularization of least-squares

$$\boldsymbol{w}^*_{\text{lasso}} = \arg\min_{\boldsymbol{w}} \underbrace{\sum_{i=1}^{N}\left(y_i - w_0 - \sum_{i=1}^{d} x_{i,j}w_j\right)^2}_{\text{RSS}} + \lambda \underbrace{\sum_{j=1}^{d}|w_j|^q}_{\text{reg. function}}$$

- $q = 1$: Lasso
- $q = 2$: Ridge regression
- $q = 0$: Subset selection

Regularization: allows complex models to be trained on data sets of limited size without severe overfitting.