OLS Coefficient Analysis

DS-GA 1013 / MATH-GA 2824 Mathematical Tools for Data Science

https://cims.nyu.edu/~cfgranda/pages/MTDS_spring20/index.html

Carlos Fernandez-Granda

Prerequisites

Linear algebra (orthogonal matrices, singular-value decomposition)

Ordinary least squares

Goal

Analyze coefficients of OLS estimator

Singular-value decomposition

Every $A \in R^{m \times k}$, $m \geq k$, has a singular-value decomposition (SVD)

$$
\begin{aligned}
A &= \begin{bmatrix} u_1 & u_2 & \cdots & u_k \end{bmatrix} \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & s_k \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \cdots & v_k \end{bmatrix}^T \\
&= USV^T
\end{aligned}
$$

The singular values $s_1 \geq s_2 \geq \cdots \geq s_k$ are nonnegative

The left singular vectors $u_1, u_2, \ldots u_k \in \mathbb{R}^m$ are orthonormal

The right singular vectors $v_1, v_2, \ldots v_k \in \mathbb{R}^k$ are orthonormal

Linear maps

The SVD decomposes the action of a matrix $A \in \mathbb{R}^{m \times k}$ on a vector $w \in \mathbb{R}^k$ into:

1. Rotation

$$V^T w = \sum_{i=1}^{k} \langle v_i, w \rangle e_i$$

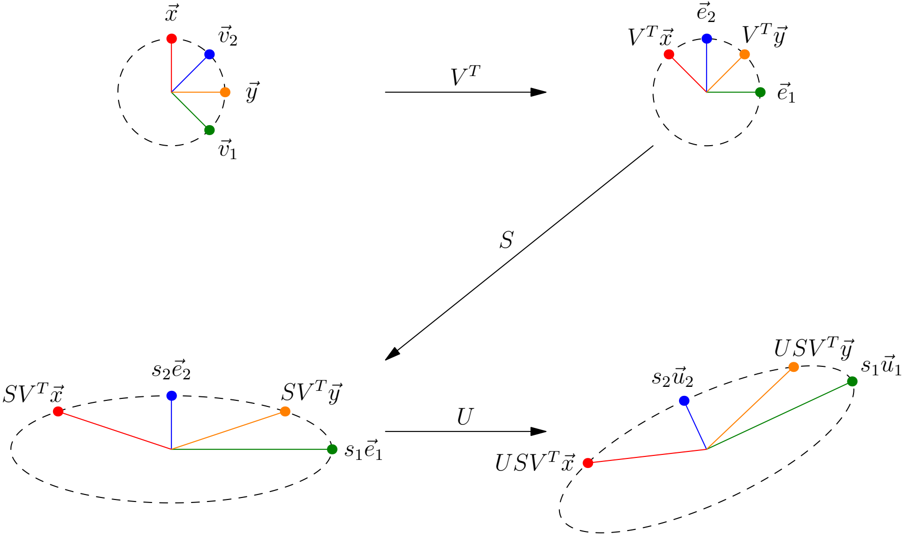
2. Scaling

$$SV^T w = \sum_{i=1}^{k} s_i \langle v_i, w \rangle e_i$$

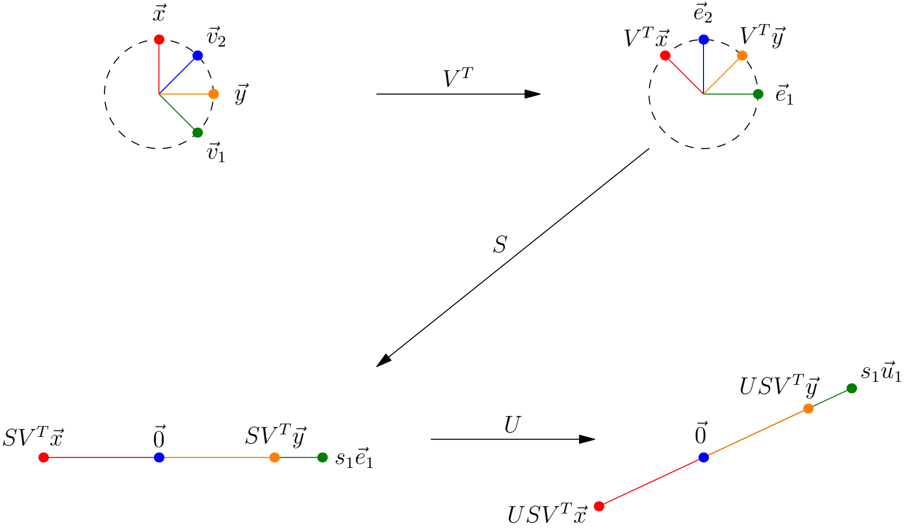
3. Rotation

$$USV^T w = \sum_{i=1}^{k} s_i \langle v_i, w \rangle u_i$$

Linear maps


$\vec{x}$
$\vec{v}_2$
$\vec{y}$
$\vec{v}_1$
V^T
$\vec{e}_2$
$V^T\vec{x}$
$V^T\vec{y}$
$\vec{e}_1$
S
$s_2\vec{e}_2$
$SV^T\vec{x}$
$SV^T\vec{y}$
$s_1\vec{e}_1$
U
$USV^T\vec{y}$
$s_1\vec{u}_1$
$s_2\vec{u}_2$
$USV^T\vec{x}$

Linear maps ($s_2 := 0$)



Regression

Goal: Estimate response (or dependent variable)

Data: Several observed variables, known as features (or covariates, or independent variables)

Data

Training data: (y_1, x_1), (y_2, x_2), ..., (y_n, x_n), where $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$

We define a response vector $y \in \mathbb{R}^n$ and a feature matrix

$$X := \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$$

OLS estimator

$$\begin{aligned}\beta_{\text{OLS}} &= \left(XX^T\right)^{-1} Xy \\ &= (US^2U^T)^{-1} USV^T y \\ &= US^{-2} U^T USV^T y \\ &= US^{-1} V^T y\end{aligned}$$

$\left(XX^T\right)^{-1} X$ is a left inverse or pseudoinverse of X^T because $\left(XX^T\right)^{-1} XX^T = I$

Additive model

Model for MSE analysis

$$\tilde{y} = \tilde{x}^T \beta_{\text{true}} + \tilde{z}$$

Problem: Ignores that estimate is obtained from training data

Model for training data

$$\tilde{y}_{\text{train}} := X^T \beta_{\text{true}} + \tilde{z}_{\text{train}}$$

- Feature matrix $X \in \mathbb{R}^{p \times n}$ is deterministic
- Coefficients $\beta_{\text{true}} \in \mathbb{R}^p$ are deterministic
- Noise $\tilde{z}_{\text{train}}$ is an n-dimensional iid Gaussian vector with zero mean and variance σ^2

Maximum likelihood

Under this model, OLS is equivalent to maximum likelihood

Assume we observe y_{train}

$$\mathcal{L}_{y_{\text{train}}}(\beta) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}\left|\left|y_{\text{train}} - X^T\beta\right|\right|_2^2\right)$$

$$\begin{aligned}
\beta_{\text{ML}} &= \arg\max_{\beta} \mathcal{L}_{y_{\text{train}}}(\beta) \\
&= \arg\max_{\beta} \log \mathcal{L}_{y_{\text{train}}}(\beta) \\
&= \arg\min_{\beta} \left|\left|y_{\text{train}} - X^T\beta\right|\right|_2^2
\end{aligned}$$

Decomposition of OLS cost function

$$\arg\min_{\beta} \|\tilde{y}_{\text{train}} - X^T\beta\|_2^2$$

$$= \arg\min_{\beta} \|\tilde{z}_{\text{train}} - X^T(\beta - \beta_{\text{true}})\|_2^2$$

$$= \arg\min_{\beta} (\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) - 2\tilde{z}_{\text{train}}^T X^T(\beta - \beta_{\text{true}}) + \tilde{z}_{\text{train}}^T \tilde{z}_{\text{train}}$$

$$= \arg\min_{\beta} (\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) - 2\tilde{z}_{\text{train}}^T X^T\beta$$

Quadratic component $(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}})$

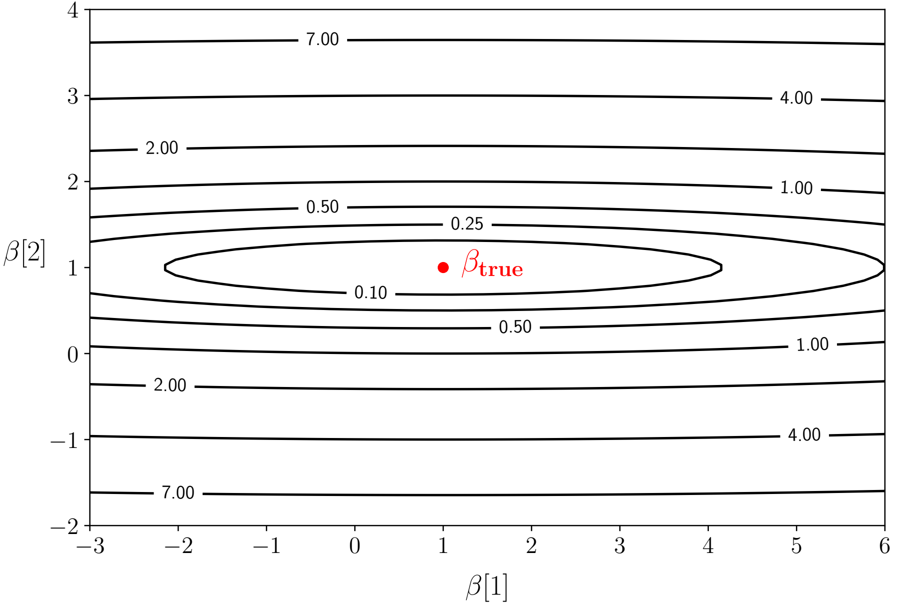
Contour lines are ellipsoids centered at β_{true}

$$
\begin{aligned}
c^2 = (\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) &= (\beta - \beta_{\text{true}})^T U S^2 U^T (\beta - \beta_{\text{true}}) \\
&= \sum_{i=1}^{p} {s_i}^2 ({u_i}^T (\beta - \beta_{\text{true}}))^2
\end{aligned}
$$

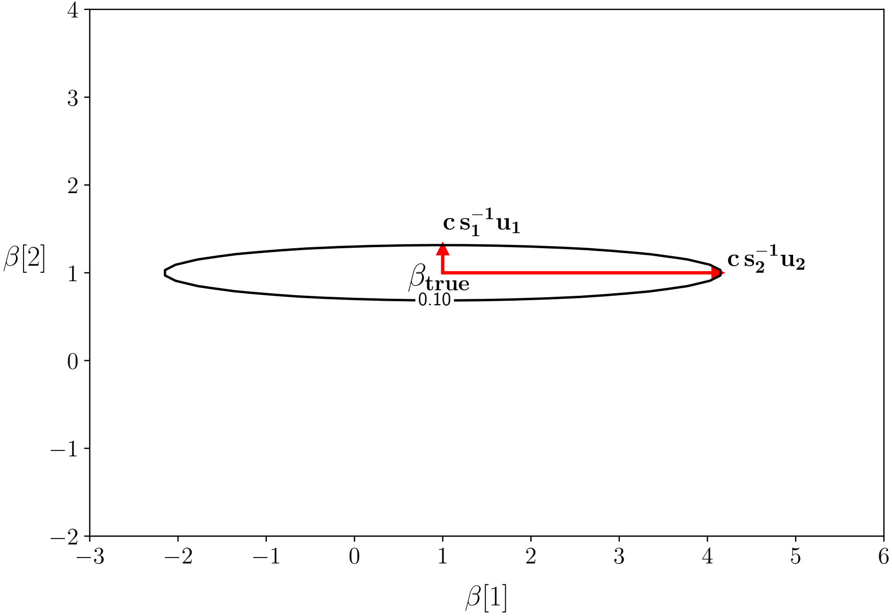
Axes of the ellipsoid are the left singular vectors

Stretch in direction of each axis is determined by singular values

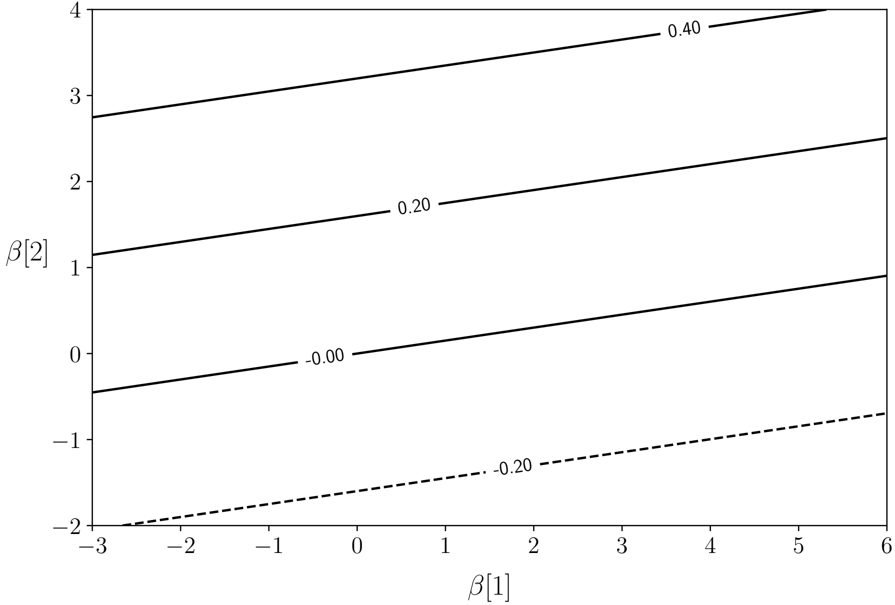
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}})$$



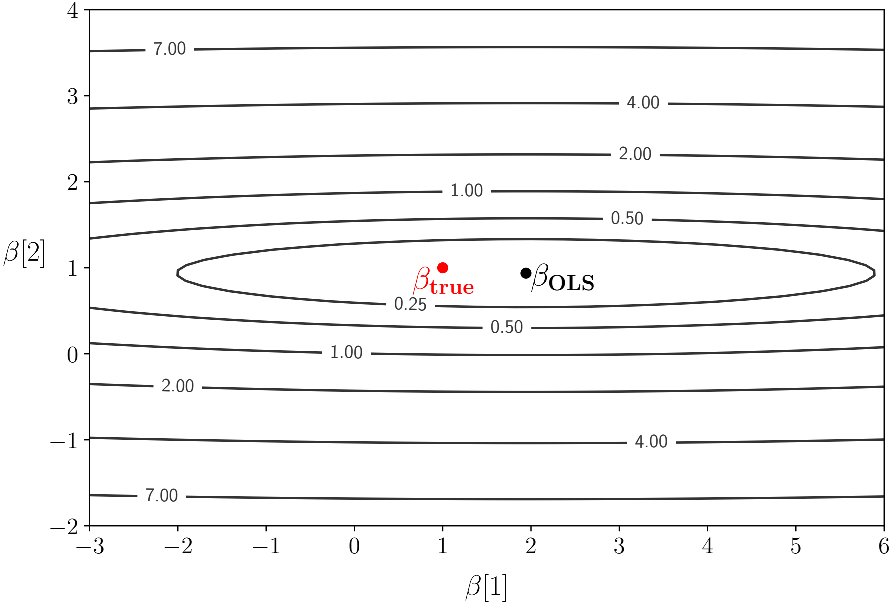
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}})$$



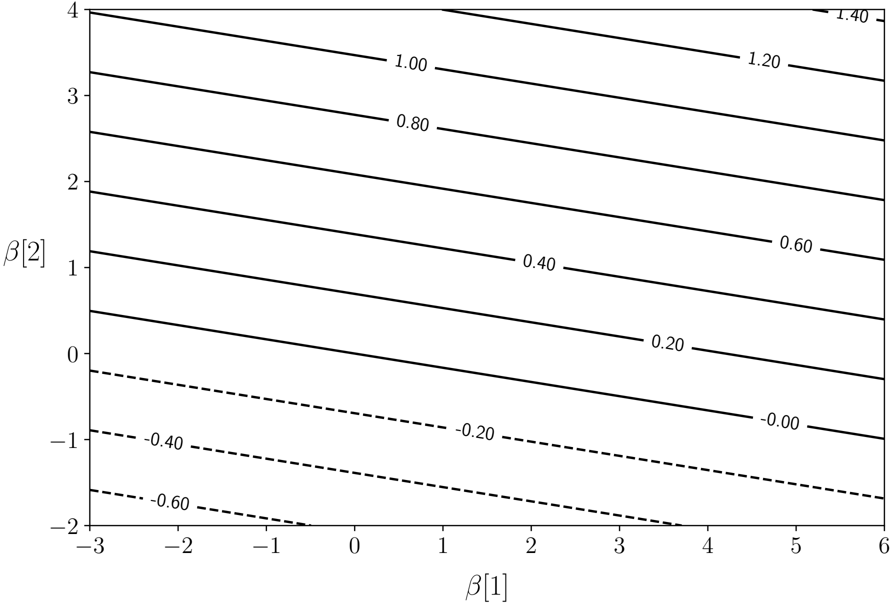
$-2\tilde{z}_{\text{train}}^T X^T \beta$



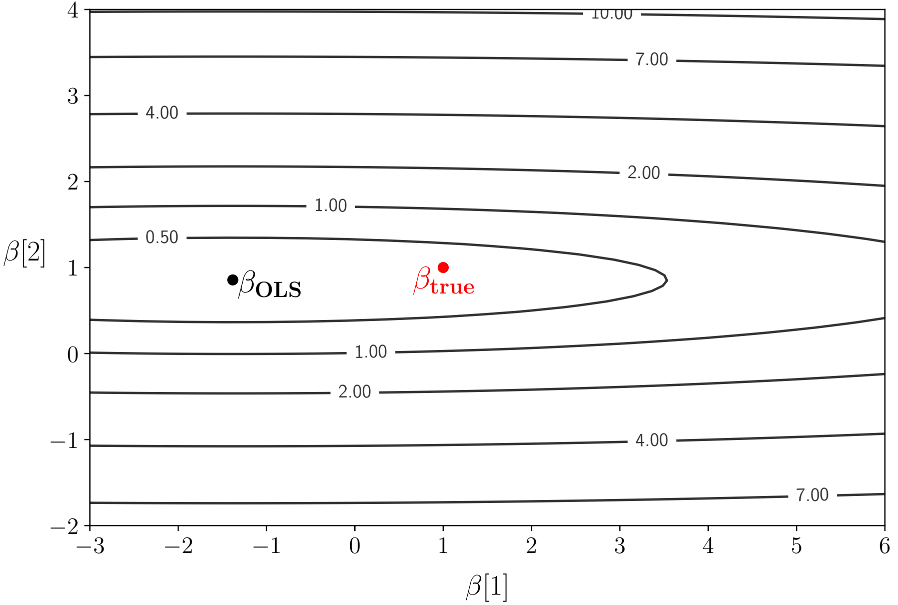
$$(\beta - \beta_{\text{true}})^T XX^T (\beta - \beta_{\text{true}}) - 2\tilde{z}_{\text{train}}^T X^T \beta$$



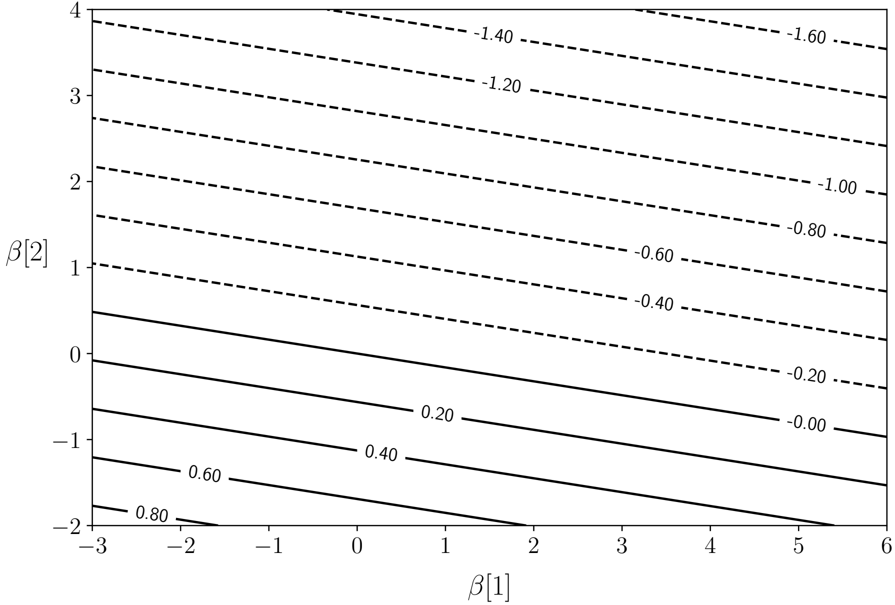
$-2\tilde{z}_{\text{train}}^T X^T \beta$



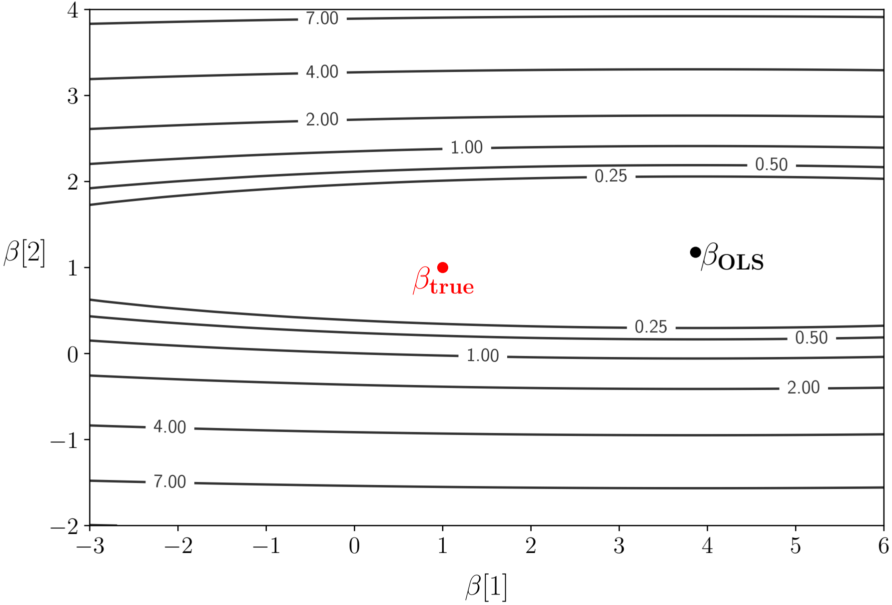
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) - 2\tilde{z}_{\text{train}}^T X^T \beta$$



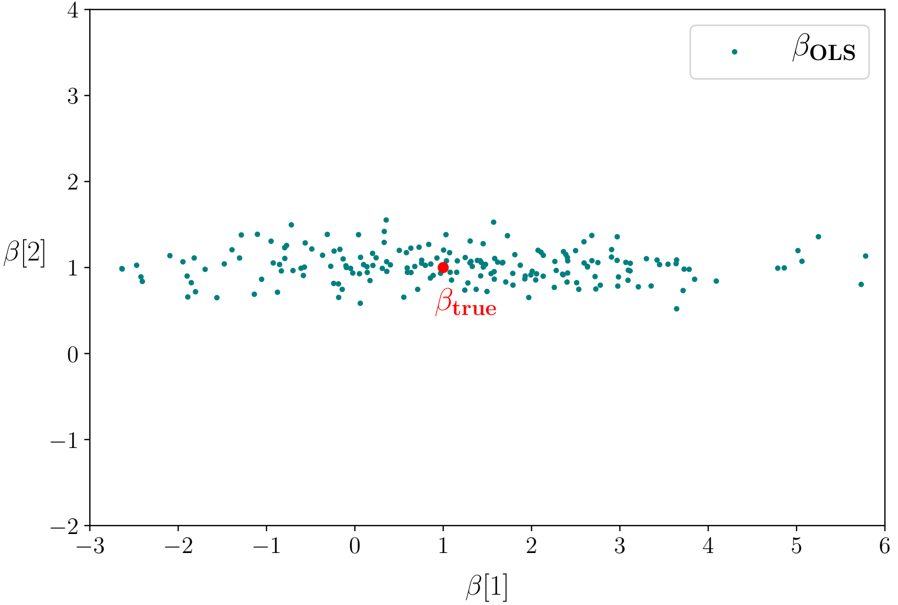
$-2\tilde{z}_{\text{train}}^T X^T \beta$



$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) - 2\tilde{z}_{\text{train}}^T X^T \beta$$



Minima for 200 realizations



Minima

$$
\begin{aligned}
\beta_{\text{OLS}} &= (XX^T)^{-1} X \tilde{y}_{\text{train}} \\
&= (XX^T)^{-1} XX^T \beta_{\text{true}} + (XX^T)^{-1} X \tilde{z}_{\text{train}} \\
&= \beta_{\text{true}} + (XX^T)^{-1} X \tilde{z}_{\text{train}} \\
&= \beta_{\text{true}} + US^{-1} V^T \tilde{z}_{\text{train}}
\end{aligned}
$$

Distribution?

Linear transformations of Gaussian random vectors

Let $\tilde{x}$ be a Gaussian random vector of dimension d with mean μ and covariance matrix Σ

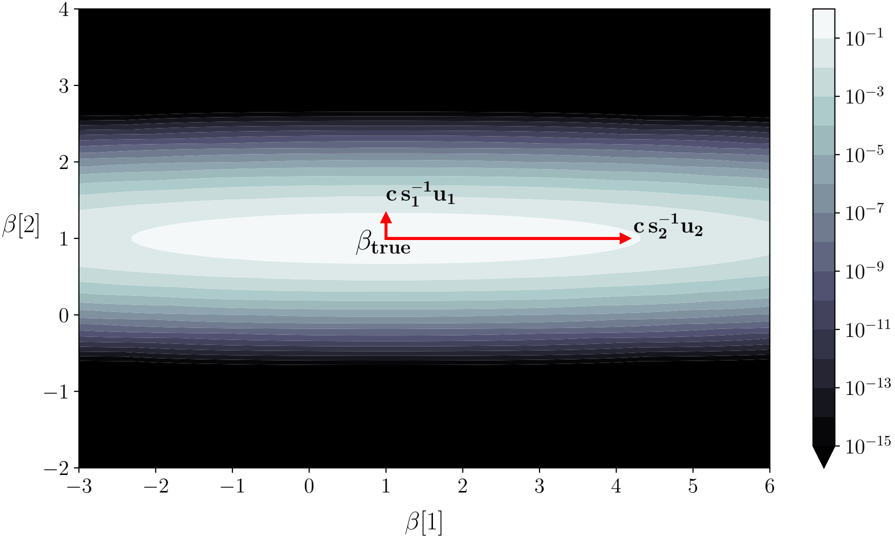
For any matrix $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^{m}$ $\tilde{y} = A\tilde{x} + b$ is Gaussian with mean $A\mu + b$ and covariance matrix $A\Sigma A^T$ (as long as it is full rank)

Minima

$$\begin{aligned}
\beta_{\text{OLS}} &= (XX^T)^{-1} X \tilde{y}_{\text{train}} \\
&= (XX^T)^{-1} XX^T \beta_{\text{true}} + (XX^T)^{-1} X \tilde{z}_{\text{train}} \\
&= \beta_{\text{true}} + (XX^T)^{-1} X \tilde{z}_{\text{train}} \\
&= \beta_{\text{true}} + US^{-1} V^T \tilde{z}_{\text{train}}
\end{aligned}$$

Distribution? Gaussian with mean β_{true} and covariance matrix $\sigma^2 US^{-2} U^T$

Minima



$\mathbf{c\,s_1^{-1}u_1}$
$\mathbf{c\,s_2^{-1}u_2}$
$\beta_{\mathbf{true}}$
$\beta[2]$
$\beta[1]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-15}

Conclusion

Coefficient error of OLS estimator depends on singular vectors and singular values of feature matrix

If singular values are small, error explodes!