Linear Regression: Coefficient Error

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at `https://www.ps4ds.net`

Regression

Goal: Estimate response from features

For example, temperature in Versailles (Kentucky) from temperatures at 133 other locations

Linear regression

Linear minimum MSE estimator of response $\tilde{y}$ given features $\tilde{x}$

$$\ell_{\mathrm{MMSE}}(\tilde{x}) = \Sigma_{\tilde{x}\tilde{y}}^{T} \Sigma_{\tilde{x}}^{-1} \left(\tilde{x} - \mu_{\tilde{x}} \right) + \mu_{\tilde{y}}$$

Key question: Do we recover the *correct* linear relationship?

Linear response with additive noise

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

Noise $\tilde{z}$ is independent from the features $\tilde{x}$

For simplicity, everything is centered to have zero mean

Linear MMSE estimator

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

$$\begin{aligned}
\beta_{\text{MMSE}} &= \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}} \\
&= \beta_{\text{true}}
\end{aligned}$$

$$\begin{aligned}
\Sigma_{\tilde{x}\tilde{y}} &= \mathrm{E}\left[\tilde{x}\tilde{y}\right] \\
&= \mathrm{E}\left[\tilde{x}\left(\tilde{x}^T \beta_{\text{true}} + \tilde{z}\right)\right] \\
&= \mathrm{E}\left[\tilde{x}\tilde{x}^T\right] \beta_{\text{true}} + \mathrm{E}\left[\tilde{x}\tilde{z}\right] \\
&= \Sigma_{\tilde{x}} \beta_{\text{true}} + \mathrm{E}\left[\tilde{x}\right] \mathrm{E}\left[\tilde{z}\right] \\
&= \Sigma_{\tilde{x}} \beta_{\text{true}}
\end{aligned}$$

End of story?

No! In practice, we compute linear models from data

Linear regression

Linear minimum MSE estimator of response $\tilde{y}$ given features $\tilde{x}$

$$\ell_{\text{MMSE}}(\tilde{x}) = \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} (\tilde{x} - \mu_{\tilde{x}}) + \mu_{\tilde{y}}$$

Ordinary-least-squares (OLS) estimator from dataset
(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)

$$\ell_{\text{OLS}}(x_i) = \Sigma_{XY}^T \Sigma_X^{-1} (x_i - m(X)) + m(Y)$$

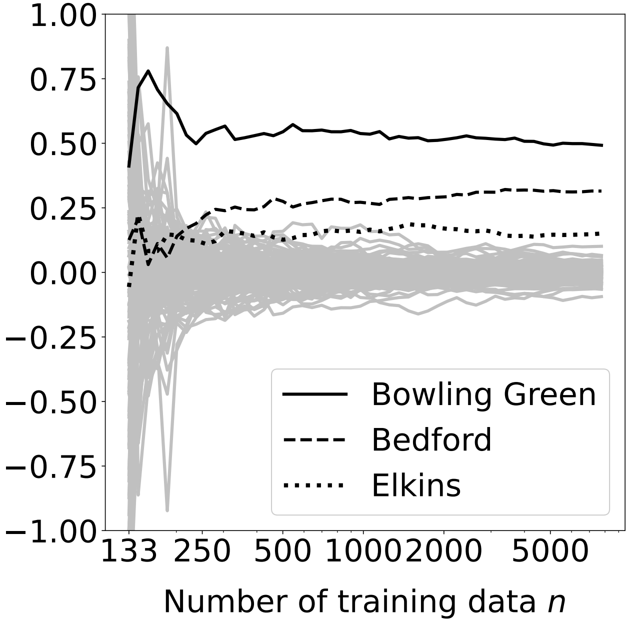
Temperature prediction

Response: Temperature in Versailles (Kentucky)

Features: Temperatures at 133 other locations

OLS coefficients (large n)

OLS coefficients



Linear response with additive noise

$$y_{\text{train}} := X_{\text{train}} \beta_{\text{true}} + z_{\text{train}}$$

$$X_{\text{train}} := \begin{bmatrix} x_1^T \\ x_2^T \\ \cdots \\ x_n^T \end{bmatrix}$$

For simplicity, everything is centered to have zero mean

OLS coefficients

$$\ell_{\text{OLS}}(x_i) = \beta_{\text{OLS}}^T x_i$$

$$\begin{aligned}\beta_{\text{OLS}} &= \Sigma_X^{-1}\Sigma_{XY} \\ &= \beta_{\text{true}} + \Sigma_X^{-1}\Sigma_{XZ}\end{aligned}$$

$$\Sigma_{XY} = \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i$$

$$\begin{aligned}\Sigma_{XY} &= \frac{1}{n-1}\sum_{i=1}^{n} x_i (x_i^T \beta_{\text{true}} + z_{\text{train}}[i]) \\ &= \left(\frac{1}{n-1}\sum_{i=1}^{n} x_i x_i^T\right)\beta_{\text{true}} + \frac{1}{n-1}\sum_{i=1}^{n} x_i z_{\text{train}}[i] \\ &= \Sigma_X \beta_{\text{true}} + \Sigma_{XZ}\end{aligned}$$

Example with independent noise samples

$$\underbrace{\begin{bmatrix} 0.33 \\ 0.91 \\ -1.51 \\ -0.10 \end{bmatrix}}_{y_{\text{train}}} := \underbrace{\begin{bmatrix} 0.46 & 0.44 \\ 0.97 & 1.03 \\ -1.52 & -1.51 \\ 0.09 & 0.04 \end{bmatrix}}_{X_{\text{train}}} \underbrace{\begin{bmatrix} 0.75 \\ 0.25 \end{bmatrix}}_{\beta_{\text{true}}} + \underbrace{\begin{bmatrix} -0.13 \\ -0.08 \\ 0.01 \\ -0.18 \end{bmatrix}}_{z_{\text{train}}}$$

$$\Sigma_{XZ} = \begin{bmatrix} -0.055 \\ -0.053 \end{bmatrix}$$

OLS coefficients

$$
\begin{aligned}
&\beta_{\text{OLS}} \\
&= \beta_{\text{true}} + \Sigma_X^{-1}\Sigma_{XZ} \\
&= \beta_{\text{true}} + U\Lambda^{-1}U^T\Sigma_{XZ} \\
&= \begin{bmatrix} 0.75 \\ 0.25 \end{bmatrix} + \begin{bmatrix} 0.70 & -0.71 \\ 0.71 & 0.70 \end{bmatrix} \begin{bmatrix} 0.43 & 0 \\ 0 & 1033 \end{bmatrix} \begin{bmatrix} 0.70 & 0.71 \\ -0.71 & 0.70 \end{bmatrix} \begin{bmatrix} -0.055 \\ -0.053 \end{bmatrix} \\
&= \begin{bmatrix} -0.71 \\ 1.65 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_X &= \begin{bmatrix} 1.15 & 1.16 \\ 1.16 & 1.17 \end{bmatrix} \\
&= \underbrace{\begin{bmatrix} 0.70 & -0.71 \\ 0.71 & 0.70 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 2.33 & 0 \\ 0 & 9.68 \cdot 10^{-4} \end{bmatrix}}_{\Lambda} \underbrace{\begin{bmatrix} 0.70 & 0.71 \\ -0.71 & 0.70 \end{bmatrix}}_{U^T}
\end{aligned}
$$

Why does noise amplification happen?

$$y_{\text{OLS}} := X_{\text{train}} \beta_{\text{OLS}}$$

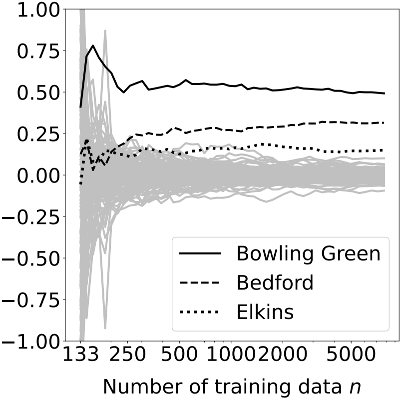
$$y_{\text{ideal}} := X_{\text{train}} \beta_{\text{true}}$$

$$\begin{aligned} ||y_{\text{OLS}} - y_{\text{train}}||_2^2 = 0.036 \quad < \quad 0.055 &= ||y_{\text{ideal}} - y_{\text{train}}||_2^2 \\ &= ||z_{\text{train}}||_2^2 \end{aligned}$$

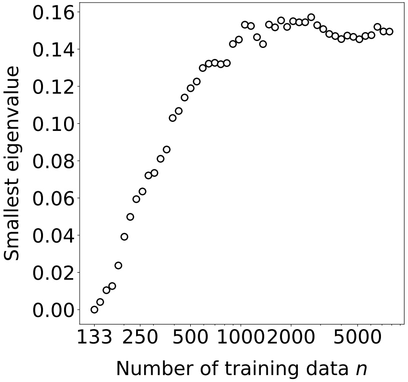
Overfitting!

Temperature prediction

OLS coefficients



Smallest eigenvalue of Σ_X



Linear response with random additive noise

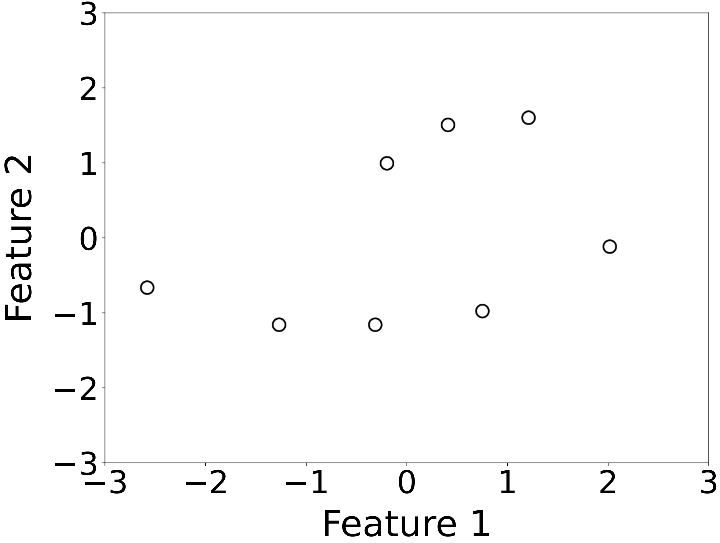
$$\tilde{y}_{\text{train}} := X_{\text{train}} \beta_{\text{true}} + \tilde{z}$$

$$X_{\text{train}} := \begin{bmatrix} x_1^T \\ x_2^T \\ \cdots \\ x_n^T \end{bmatrix}$$

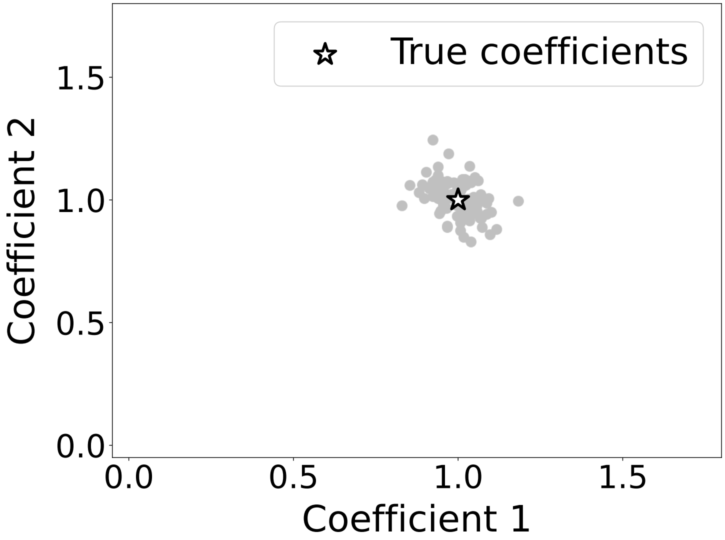
Noise $\tilde{z}$ is i.i.d. with variance σ^2 and independent from the features

For simplicity, everything is centered to have zero mean

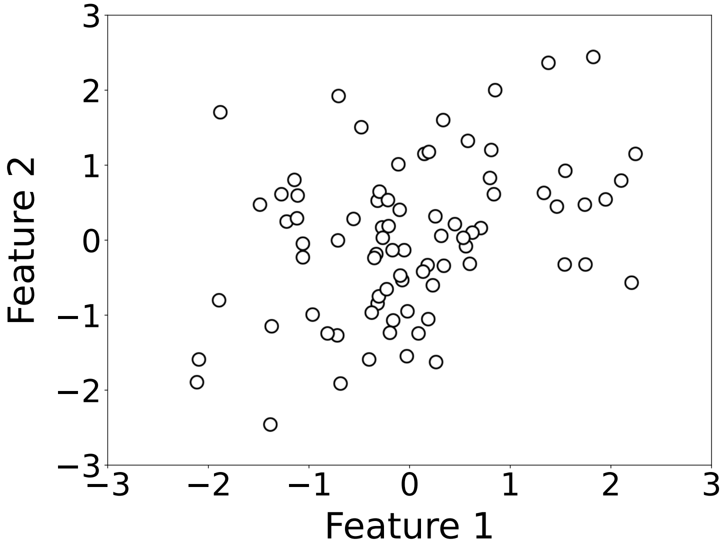
Features ($n := 8$)



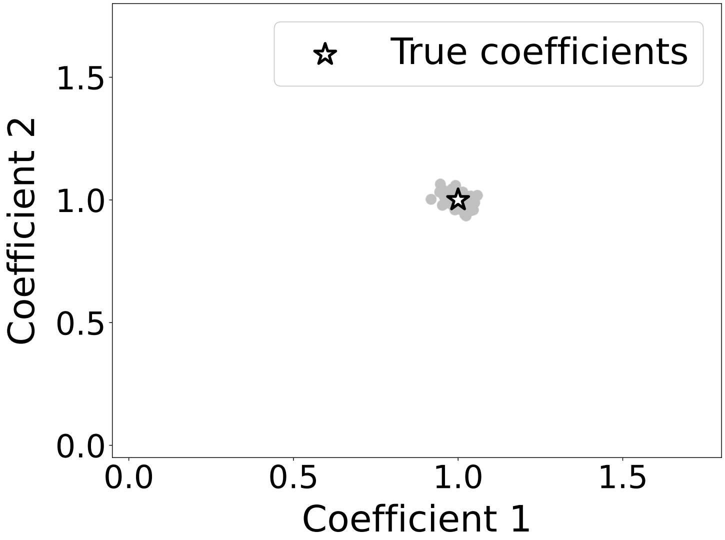
100 coefficient estimates



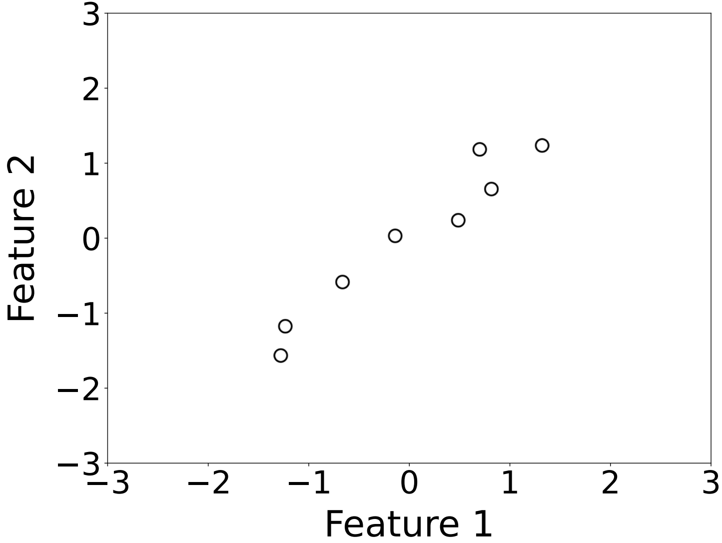
Features ($n := 80$)



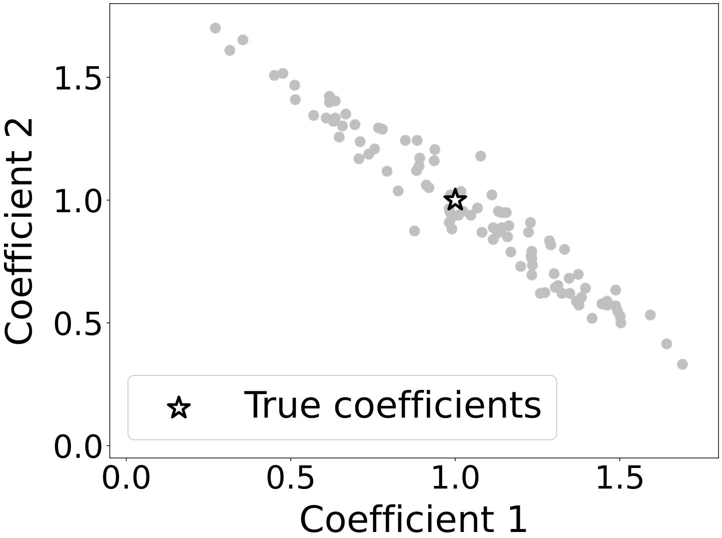
100 coefficient estimates



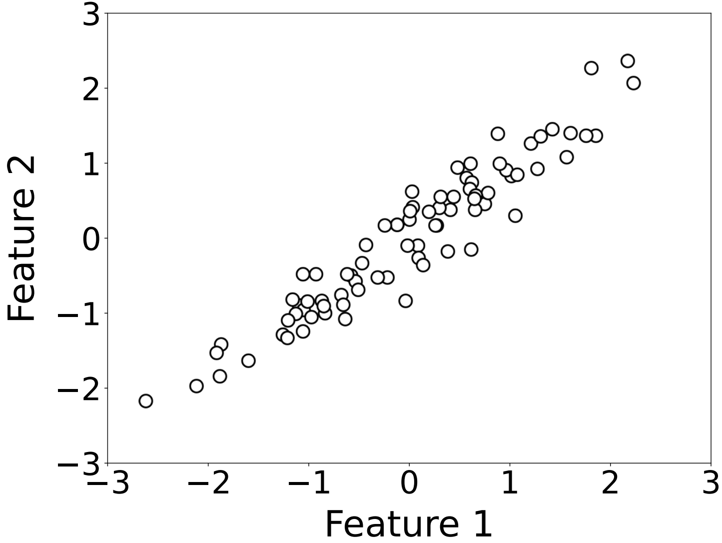
Collinear features ($n := 8$)



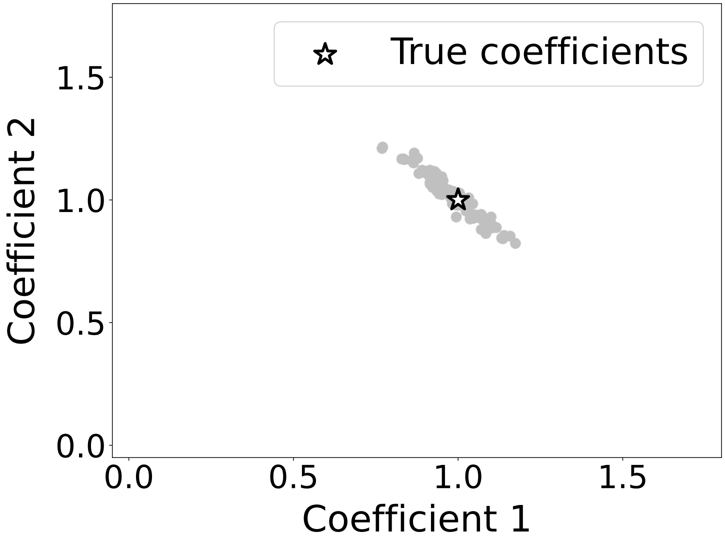
100 coefficient estimates



Collinear features ($n := 80$)



100 coefficient estimates



Empirical observations

- OLS coefficients are centered at true coefficients
- Variance decreases as number of training data n grows
- When features are collinear, variance is large in directions of low feature variance

OLS coefficients

$$\tilde{\beta}_{\mathsf{OLS}} = \beta_{\mathsf{true}} + \Sigma_X^{-1} \widetilde{\Sigma}_{XZ}$$

$$\widetilde{\Sigma}_{XZ} := \frac{1}{n-1} \sum_{i=1}^{n} x_i \tilde{z}_i$$

$$\begin{aligned} \mathrm{E}\left[\tilde{\beta}_{\mathsf{OLS}}\right] &= \beta_{\mathsf{true}} + \Sigma_X^{-1} \mathrm{E}\left[\widetilde{\Sigma}_{XZ}\right] \\ &= \beta_{\mathsf{true}} + \Sigma_X^{-1} \frac{1}{n-1} \sum_{i=1}^{n} x_i \mathrm{E}\left[\tilde{z}_i\right] \\ &= \beta_{\mathsf{true}} \qquad \text{Unbiased!} \end{aligned}$$

Covariance matrix of coefficients

$$\tilde{\beta}_{\mathsf{OLS}} = \beta_{\mathsf{true}} + \Sigma_X^{-1}\widetilde{\Sigma}_{XZ}$$

$$\mathsf{ct}\left(\tilde{\beta}_{\mathsf{OLS}}\right) = \Sigma_X^{-1}\widetilde{\Sigma}_{XZ}$$

$$\begin{aligned}
\Sigma_{\tilde{\beta}_{\mathsf{OLS}}} &= \mathrm{E}\left[\mathsf{ct}\left(\tilde{\beta}_{\mathsf{OLS}}\right)\mathsf{ct}\left(\tilde{\beta}_{\mathsf{OLS}}\right)^T\right] \\
&= \mathrm{E}\left[\Sigma_X^{-1}\widetilde{\Sigma}_{XZ}\widetilde{\Sigma}_{XZ}^T\Sigma_X^{-1}\right] \\
&= \Sigma_X^{-1}\mathrm{E}\left[\widetilde{\Sigma}_{XZ}\widetilde{\Sigma}_{XZ}^T\right]\Sigma_X^{-1}
\end{aligned}$$

Covariance matrix of coefficients

$$\begin{aligned}
\Sigma_{\tilde{\beta}_{\text{OLS}}} &= \Sigma_X^{-1} \mathrm{E}\left[\widetilde{\Sigma}_{XZ} \widetilde{\Sigma}_{XZ}^T\right] \Sigma_X^{-1} \\
&= \frac{\sigma^2}{n-1} \Sigma_X^{-1} \Sigma_X \Sigma_X^{-1} = \frac{\sigma^2}{n-1} \Sigma_X^{-1} \qquad \propto \quad \frac{1}{n}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left[\widetilde{\Sigma}_{XZ} \widetilde{\Sigma}_{XZ}^T\right] &= \mathrm{E}\left[\frac{1}{n-1} \sum_{i=1}^{n} x_i \tilde{z}_i \frac{1}{n-1} \sum_{j=1}^{n} \tilde{z}_j x_j^T\right] \\
&= \frac{1}{(n-1)^2} \sum_{i=1}^{n} \sum_{j=1}^{n} x_i \mathrm{E}\left[\tilde{z}_i \tilde{z}_j\right] x_j^T \\
&= \frac{\sigma^2}{(n-1)^2} \sum_{i=1}^{n} x_i x_i^T \\
&= \frac{\sigma^2}{n-1} \Sigma_X
\end{aligned}$$

PCA perspective

Feature sample covariance matrix:

$$\Sigma_X = \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix} \begin{bmatrix} \xi_1 & 0 & \cdots & 0 \\ 0 & \xi_2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \xi_d \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix}^T$$

$$\begin{aligned} \Sigma_{\tilde{\beta}_{\text{OLS}}} &= \frac{\sigma^2}{n-1} \Sigma_X^{-1} \\ &= \frac{\sigma^2}{n-1} \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix} \begin{bmatrix} 1/\xi_1 & 0 & \cdots & 0 \\ 0 & 1/\xi_2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & 1/\xi_d \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix}^T \end{aligned}$$

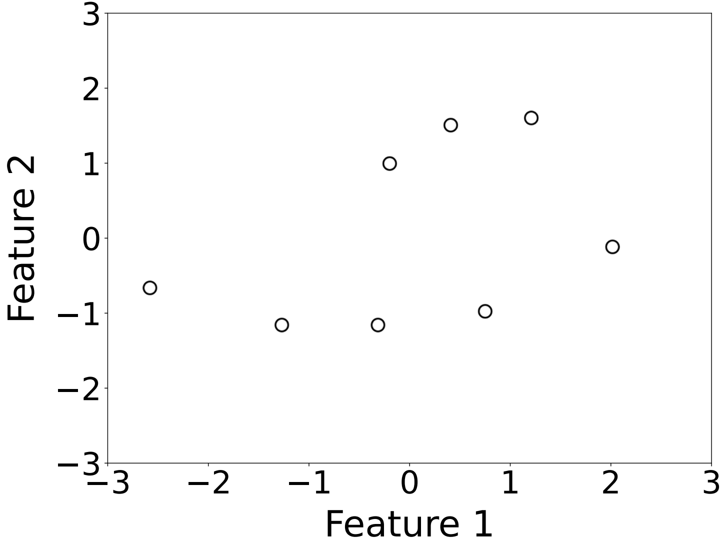
Variance in the jth principal direction of the features:

$$\operatorname{Var}\left[u_j^T \tilde{\beta}_{\text{OLS}}\right] = \frac{\sigma^2}{(n-1)\,\xi_j}$$

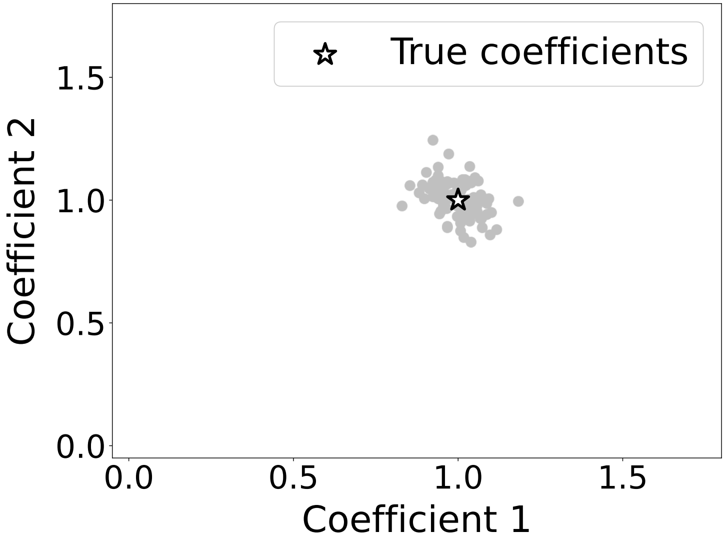
Theoretical conclusions

- OLS coefficients are unbiased (centered at true coefficients)
- The estimator is consistent (error tends to zero as number of training data n grows)
- Coefficient variance is large in feature principal directions with low variance

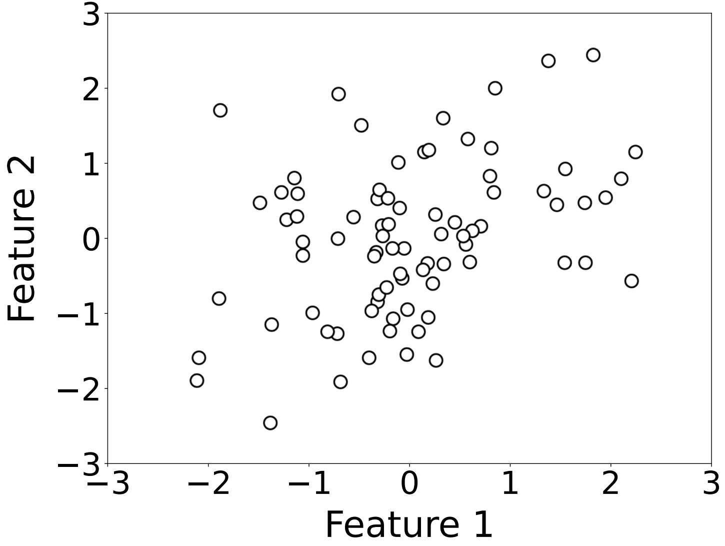
Non-collinear features ($n := 8$)



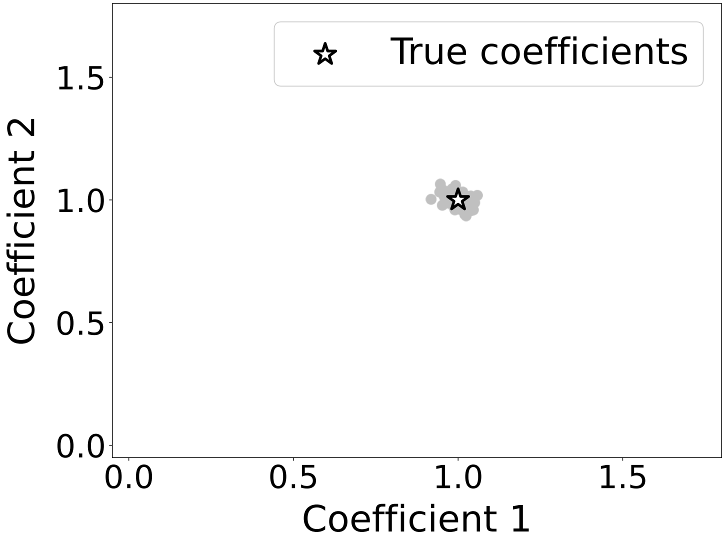
OLS coefficients



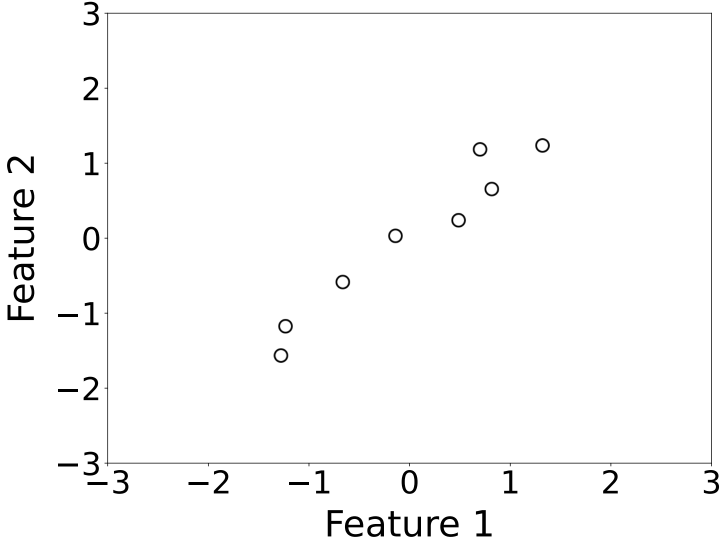
Non-collinear features ($n := 80$)



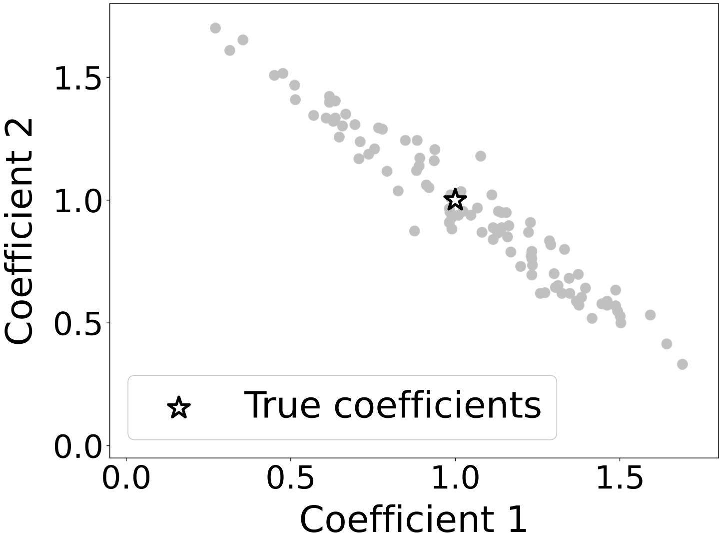
OLS coefficients



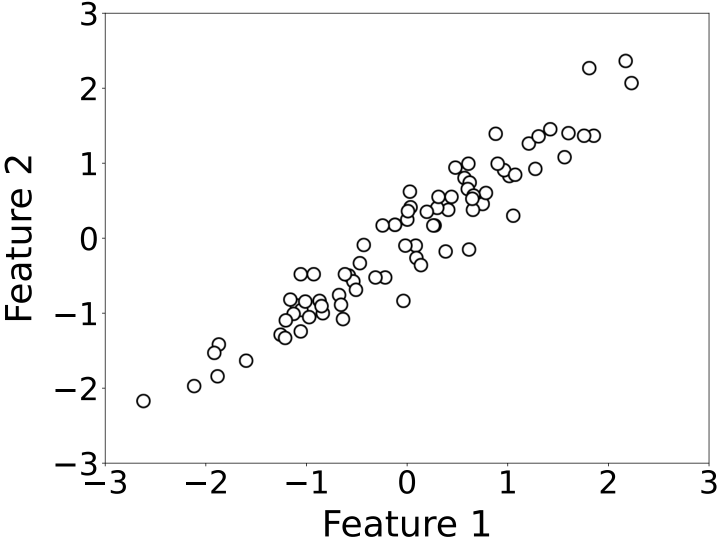
Collinear features ($n := 8$)



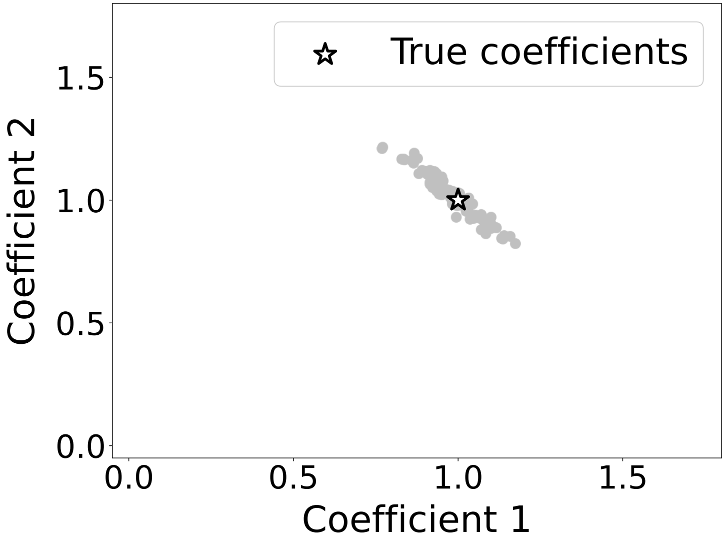
OLS coefficients



Non-collinear features ($n := 80$)



OLS coefficients



What have we learned?

OLS coefficients recover linear structure, if there's enough data

OLS coefficient estimator is unbiased, but can have large variance due to feature collinearity