On PCA and Linear Regression

CS189/289A: Introduction to Machine Learning

Stella Yu

UC Berkeley

25 September 2018

Outline: Connections of PCA to Linear Regression

1. Singular Value Views of Linear Regression
2. Dual Space Views and Kernel Functions
3. Total Least Squares for Linear Regression

LS Regression from One Space to Another

- X contains n points in d-dimensional space
  y contains corresponding responses:

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}, \qquad y = \begin{bmatrix} y1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{1}$$

- Both assumed zero mean.

- Linear regression model w maps x_i to y_i:

$$y_i = x_i' w, \qquad y = Xw \tag{2}$$

LS Regression Solution: Primal

- Singular value decomposition of data matrix X:

$$X = \underbrace{U_{n\times n}}_{\text{sample space basis}} \quad \underbrace{S_{n\times d}}_{\text{feature to sample space scaling}} \quad \underbrace{V_{d\times d}}_{\text{feature space basis}}' \tag{3}$$

- $\forall s_i = S_{ii} > 0$, there is a column vector pair (U_i, V_i).
- Least-square solutions take the following form:

$$w = \sum_{i=1}^{n} V_i \cdot \rho(s_i) \cdot (U_i' y) \tag{4}$$

  $\rho(s)$ is a weight function of singular values.

- X's two eigenspaces decide the structure of the solution, no matter what y is.
- y is simply projected onto the sample space of X.

Singular Value View of Pseudo-Inverse

$$
\begin{aligned}
Xw &= y && (5)\\
X = USV' &= U\begin{bmatrix} S_+ & \\ & 0 \end{bmatrix} V' && (6)\\
USV'w &= y && (7)\\
SV'w &= U'y && (8)\\
w &= VS^{\dagger}U'y && (9)\\
&= V\begin{bmatrix} S_+^{-1} & \\ & 0 \end{bmatrix} U'y && (10)
\end{aligned}
$$

$$
w = \sum_{i=1}^{n} V_i \cdot \rho_{\mathrm{pinv}}(s_i) \cdot (U_i'y) \tag{11}
$$

$$
\rho_{\mathrm{pinv}}(s) = \begin{cases} \frac{1}{s}, & s > 0 \\ 0, & s = 0 \end{cases} \tag{12}
$$

Singular Value View of PCA for LS

$$Xw = y \tag{13}$$

$$X = USV' \approx U \begin{bmatrix} S_k & \\ & 0 \end{bmatrix} V' \tag{14}$$

$$U \begin{bmatrix} S_k & \\ & 0 \end{bmatrix} V'w = y \tag{15}$$

$$w = V \begin{bmatrix} S_k^{-1} & \\ & 0 \end{bmatrix} U'y \tag{16}$$

$$w = \sum_{i=1}^{n} V_i \cdot \rho_{\text{PCA}}(s_i) \cdot (U_i' y) \tag{17}$$

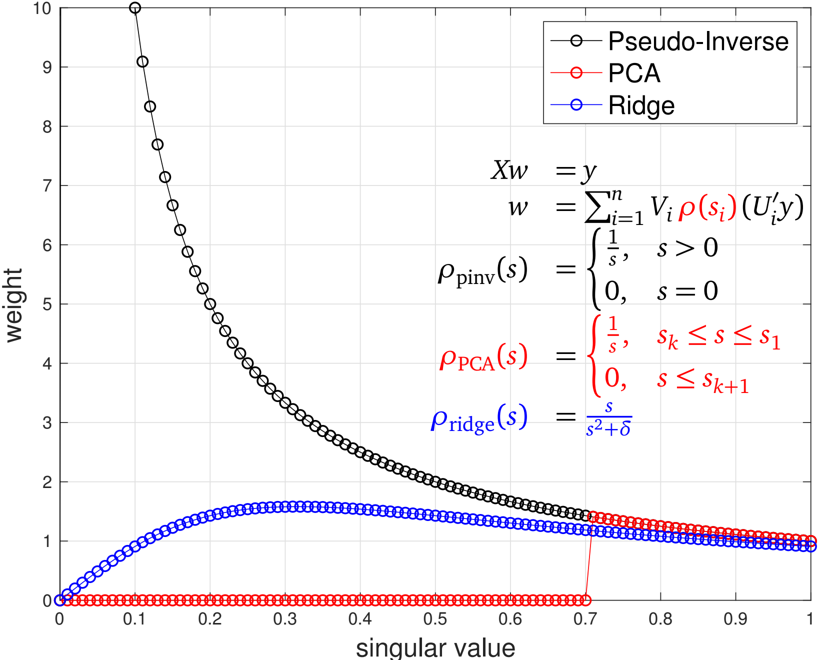
$$\rho_{\text{PCA}}(s) = \begin{cases} \frac{1}{s}, & s_k \le s \le s_1 \\ 0, & s \le s_{k+1} \end{cases} \tag{18}$$

Singular Value View of Ridge Regression

$$
\begin{aligned}
Xw &= y && (19)\\
X'Xw &= X'y && (20)\\
X = USV' &= U\begin{bmatrix} S_+ & \\ & 0 \end{bmatrix}V' && (21)\\
X'X &= VS'SV' && (22)\\
(X'X+\delta I)w &= X'y && (23)\\
V(S'S+\delta I)V'w &= VS'U'y && (24)\\
w &= V(S'S+\delta I)^{-1}S'U'y && (25)
\end{aligned}
$$

$$
\begin{aligned}
w &= \sum_{i=1}^{n} V_i \cdot \rho_{\text{ridge}}(s_i) \cdot (U_i'y) && (26)\\
\rho_{\text{ridge}}(s) &= \frac{s}{s^2+\delta} && (27)
\end{aligned}
$$

Singular Value Weight for Pinv, PCA, Ridge



LS Regression Solution: Dual

- Least-square solutions take the following form:

$$w = \sum_{i=1}^{n} V_i \cdot \rho(s_i) \cdot (U_i' y) = V \cdot \rho(S) \cdot U' y \tag{28}$$

- w lies in the column space of V, i.e., the row space of X.
- w can thus be expressed as a combination of X's row vectors:

$$w = X' \alpha \tag{29}$$

- α lies in the column space of U and is the dual vector of w.

$$Xw = XX'\alpha = y \tag{30}$$

$$USS'U'\alpha = y \tag{31}$$

$$\alpha_{\text{pinv}} = U(SS')^{\dagger} U' y \tag{32}$$

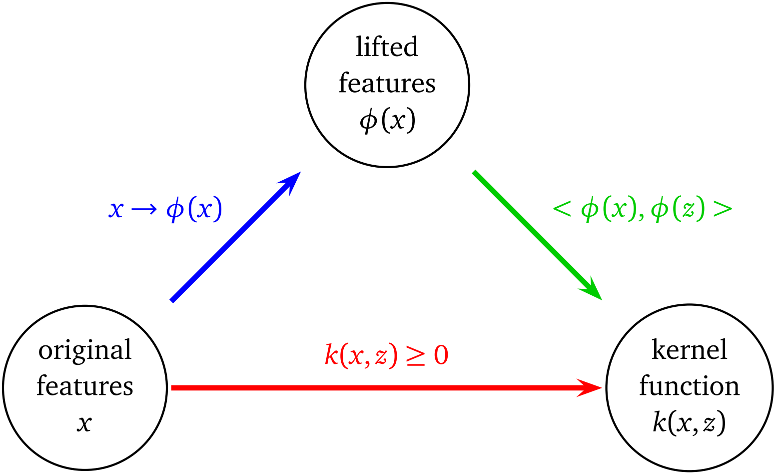
Ridge Regression Solution: Dual

$$(X'X + \delta I)w = X'y \tag{33}$$
$$(X'X + \delta I)X'\alpha = X'y \tag{34}$$
$$(X'XX' + \delta X')\alpha = X'y \tag{35}$$
$$X'(XX' + \delta I)\alpha = X'y \tag{36}$$
$$XX'(XX' + \delta I)\alpha = XX'y \tag{37}$$
$$USS'U'(USS'U' + \delta I)\alpha = USS'U'y \tag{38}$$
$$SS'(SS' + \delta I)U'\alpha = SS'U'y \tag{39}$$
$$\alpha = U \cdot (SS' + \delta I)^{-1} \cdot U'y + \alpha_0 \tag{40}$$
$$w = V \cdot (S'S + \delta I)^{-1} S' \cdot U'y \tag{41}$$

Note: The general solution w to $Aw = b$, A rank-deficient, contains one specific solution w_b and one solution w_0 in the null(A):

$$Aw = b \quad \Rightarrow \quad Aw_b = b \tag{42}$$
$$w = w_b + w_0 \qquad\qquad Aw_0 = 0 \tag{43}$$

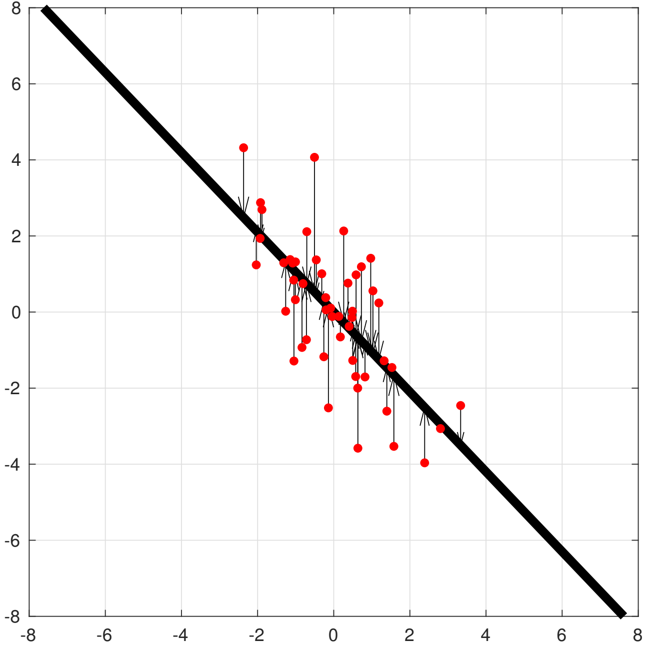
Review: Data, Feature Mapping, and Kernel



Ridge Regression Solution: Kernel

$$
\begin{aligned}
(X'X + \delta I)w &= X'y && (44)\\
X'(XX' + \delta I)\alpha &= X'y && (45)\\
XX'(XX' + \delta I)\alpha &= XX'y && (46)\\
K(K + \delta I)\alpha &= Ky && (47)\\
\alpha &= (K + \delta I)^{-1}y + \alpha_0 && (48)\\
K\alpha_0 &= 0, \quad X'\alpha_0 = 0 && (49)\\
\hat{z} &= z'w && (50)\\
&= z'X'\alpha && (51)\\
&= z'X'(K + \delta I)^{-1}y && (52)\\
\hat{z} &= \mathrm{ker}(z, X) \cdot (\mathrm{ker}(X, X) + \delta I)^{-1}y && (53)
\end{aligned}
$$

Linear Regression: LS Result



Linear Regression: LS Model

- Model: Assume noiseless data X and noisy response y

$$Y + N_Y = Xw \tag{54}$$

$$N_Y \sim \mathcal{N}(0, \sigma_Y^2) \tag{55}$$

- LS criterion: Minimize vertical projection distance

$$\min_u \varepsilon_{LS}(u) = \|y - Xu\|^2 \tag{56}$$

$$= \left\| [y, X] \begin{bmatrix} 1 \\ -u \end{bmatrix} \right\|^2 \tag{57}$$

- LS solution as regression from X to y:

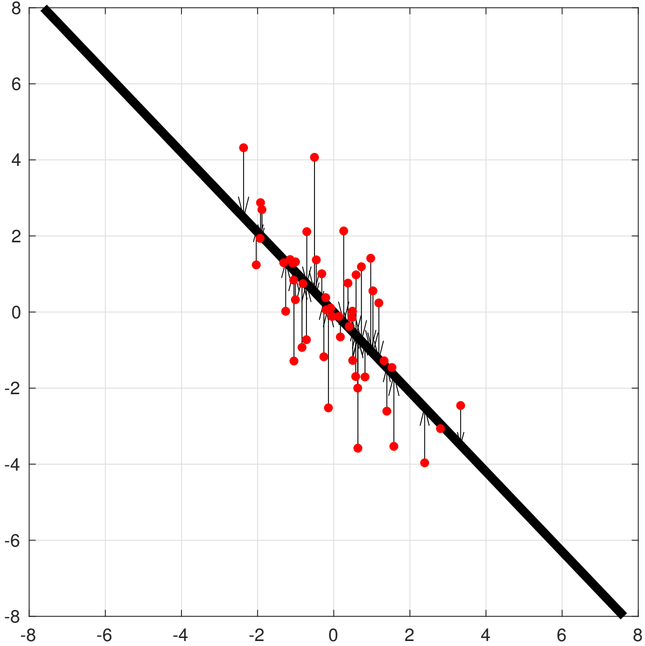
$$y = Xu \Rightarrow u = X^\dagger y \tag{58}$$

- LS solution as variance reduction in the joint (X, y) space:

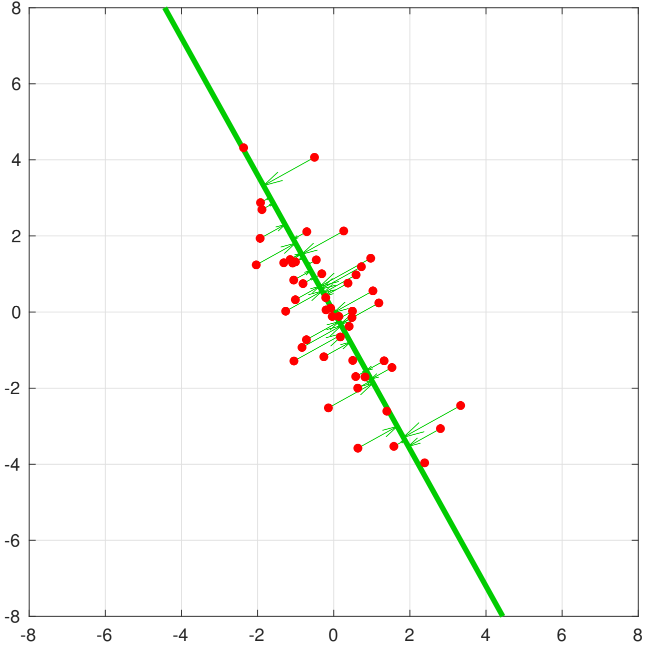
$$\begin{bmatrix} y & X \end{bmatrix} \begin{bmatrix} 1 \\ u \end{bmatrix} = USV' \begin{bmatrix} 1 \\ u \end{bmatrix} = 0 \Rightarrow SV' \begin{bmatrix} 1 \\ u \end{bmatrix} = 0 \Rightarrow \begin{bmatrix} 1 \\ u \end{bmatrix} \propto V_{\text{last}} \tag{59}$$

- In LS, u is the normal direction of the line.

Linear Regression: LS Result



Linear Regression: Total Least Squares Result



Linear Regression: TLS Model

- Model: Assume noise in both data X and response y

$$Y + N_Y = (X + N_X)w \tag{60}$$

$$N_X \sim \mathcal{N}(0, \sigma_X^2) \tag{61}$$

$$N_Y \sim \mathcal{N}(0, \sigma_Y^2) \tag{62}$$

- TLS criterion: Minimize orthogonal projection distance

$$\min_u \varepsilon_{TLS}(u) = \|[y, X] - [y, X]uu'\|^2 \tag{63}$$

$$= \|[y, X](I - uu')\|^2 \tag{64}$$

$$= \min_u \varepsilon_{PCA\_Err}(u), \quad u'u = 1 \tag{65}$$

$$= \text{constant} - \max_u \varepsilon_{PCA\_Var}(u), \quad u'u = 1 \tag{66}$$

$$= \text{constant} - \max_u \|[y, X]u\|^2, \quad u'u = 1 \tag{67}$$

- TLS = PCA solution in the joint (X, y) space:

$$\begin{bmatrix} y & X \end{bmatrix} u = USV'u \Rightarrow u = V_1 \tag{68}$$

- In TLS, u is the direction of the line.

PCA for LS and TLS