Principal Components

Let $\bar{C} = \bar{X}^T \bar{X} = \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T$

$\bar{C}$ has eigenvalue decomposition

$$\bar{C} = \bar{V} \bar{D} \bar{V}^T \qquad \bar{D} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix}, \quad V V^T = I$$

$\lambda_i \geq 0$ because $\bar{C}$ is positive semidefinite

$$\left(\text{if } \bar{C} z = \lambda_i z, \quad \bar{z}^T \bar{C} \bar{z} \geq 0 \Rightarrow \lambda_i \geq 0 \right)$$

Recall that if $d \leq n$ and $\bar{X}$ has full rank

$$\vec{w}_{OLS} = (\bar{X}^T \bar{X})^{-1} \bar{X}^T \vec{y}$$

<u>Now:</u> $\vec{w}_{OLS} = \sum_{i=1}^{d} \frac{\alpha_i}{\lambda_i} \vec{v}_i$

where $\alpha_i = \sum_{j=1}^{n} \underbrace{y_j}_{\text{value of } \vec{x}_j} \underbrace{\langle \vec{x}_j, \vec{v}_i \rangle}_{\text{similarity of } \vec{x}_j \text{ and } \vec{v}_i}$

Decomposes $\vec{w}_{OLS}$ into d-feature directions which are a basis. Directions weighted by

λ_i : presence of signal in data

α_i : interpolated value in v_i direction.

<u>NOTE:</u> Small λ_i get amplified by $\frac{1}{\lambda_i}$!

<u>SVD:</u> $n \geq d$

$$\bar{X} = \bar{U}\,\bar{S}\,\bar{V}^T \qquad \bar{U} \;\; n \times n, \qquad \bar{V} \;\; d \times d$$

$$\bar{S} = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_d \\ & 0 & \end{bmatrix} \Big\}\, n-d$$

$$\bar{C} = \bar{X}^T \bar{X} = \bar{V}\,(\bar{S}^T \bar{S})\,\bar{V}^T \implies \sigma_i^2 = \lambda_i$$

same $\bar{V}$!

<u>Note</u>: $\vec{w}_{OLS} = \bar{C}^{-1} \bar{X}^T \vec{y}$

$$\implies \vec{w}_{OLS} = \sum_{i=1}^{d} \frac{1}{\sigma_i} \langle \vec{u}_i, \vec{y} \rangle\, \vec{v}_i$$

Again, small σ_i get amplified.

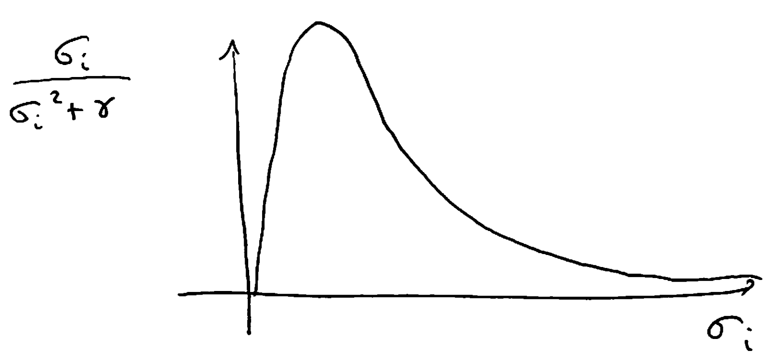
If $\sigma_i \approx$ should probably ignore it!

<u>RIDGE</u> <u>REGRESSION</u>: Pick $\gamma > 0$ and define

$$\vec{w}_{RIDGE} = \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2 + \gamma} \langle \vec{u}_i, \vec{y} \rangle\, \vec{v}_i$$

If σ_i is very small, $\dfrac{\sigma_i}{\sigma_i^2+\gamma}$ is small

If σ_i is very large, $\dfrac{\sigma_i}{\sigma_i^2+\gamma} \approx \dfrac{1}{\sigma_i}$



Underdetermined regression

If $\bar{X}$ is not full rank, ridge still works.

$$\sigma_i = 0 \implies \frac{\sigma_i}{\sigma_i^2 + \gamma} = 0$$

Ignores directions in null $(\bar{X})$.

$$\vec{w}_{PINV} = \sum_{i=1}^{d} r_i \langle \vec{u}_i, \vec{y} \rangle \vec{v}_i$$

$$r_i = \begin{cases} \dfrac{1}{\sigma_i} & \sigma_i > 0 \\ 0 & \sigma_i = 0 \end{cases}$$

limit of ridge as $\gamma \to 0$.

same as $\vec{w}_{PINV} = \bar{X}^{\dagger} \vec{y}$

pseudoinverse of $\bar{X}$ times $\bar{y}$.

Ignores null $(\bar{X})$, .

Ridge regression

Optimization View

Least squares: $\underset{w}{\text{minimize}} \ \| \bar{X} \vec{w} - \vec{y} \|^2$

if $n < d$ OR $\bar{X}$ rank deficient, there are infinitely many solutions.

if $\vec{w}_*$ is a minimizer, so is $\vec{w}_* + \vec{u}$ for $u \in \text{null}(\bar{X})$.

Consider: $\underset{w}{\text{minimize}} \ \| \bar{X} \vec{w} - \vec{y} \|^2 + \underbrace{\gamma \| \vec{w} \|^2}_{\text{penalty term on } \vec{w}.}$

Write $\vec{w} = \vec{w}_x + \vec{w}_\perp$ where

$\vec{w}_x \in \text{range}(\bar{X}^T) = \text{span}(\{\bar{X}_i\})$

$\vec{w}_\perp \in \text{null}(\bar{X})$

Then $\| \bar{X} \vec{w} - \vec{y} \| + \gamma \| \vec{w} \|^2$

$= \| \bar{X} \vec{w}_x - \vec{y} \|^2 + \gamma \| \vec{w}_x \|^2 + \gamma \| \vec{w}_\perp \|^2$

minimizer sets $w_\perp = 0$

Ridge solution:

$$\|\bar{X}(\vec{w}_\star + \Delta\vec{w}) - \vec{y}\|^2 + \gamma\|\vec{w}_\star + \Delta\vec{w}\|^2 \geq \|\bar{X}\vec{w}_\star - \vec{y}\|^2 + \gamma\|\vec{w}_\star\|^2 \qquad \forall\ \Delta\vec{w}$$

$$\Rightarrow\ 2\langle \bar{X}\vec{w}_\star - \vec{y},\ \bar{X}\Delta\vec{w}\rangle + \|\bar{X}\Delta\vec{w}\|^2 + 2\gamma\langle\vec{w}_\star,\ \Delta\vec{w}\rangle + \gamma\|\Delta\vec{w}\|^2 \geq 0 \qquad \forall\ \Delta\vec{w}$$

$$\Rightarrow\ \bar{X}^T(\bar{X}\vec{w}_\star - \bar{y}) + \gamma w_\star = 0$$

$$\Rightarrow\ (\bar{X}^T\bar{X} + \gamma\bar{I})\, w_\star = \bar{X}^T\bar{y}$$

$$\Rightarrow\ w_\star = (\bar{X}^T\bar{X} + \gamma\bar{I})^{-1}\bar{X}^T\bar{y}$$

$\vec{w}_{RIDGE}$ is unique.

Alternative Solutions

We observed $\vec{w}_{RIDGE} = \bar{X}^T \vec{\beta}$ for some $\vec{\beta}$

We can optimize for $\vec{\beta}$:

$$\underset{\vec{\beta}}{\text{minimize}} \quad \| \bar{X} \bar{X}^T \vec{\beta} - \vec{y} \|^2 + \gamma \| \bar{X}^T \vec{\beta} \|^2$$

Define $\bar{G} = \bar{X} \bar{X}^T$ $\quad n \times n \quad$ $G_{ij} = \langle x_i, x_j \rangle$

Then $\| \bar{G} \vec{\beta} - \vec{y} \|^2 + \gamma \vec{\beta}^T \bar{G} \vec{\beta}$ is

minimized if

$$\bar{G} (\bar{G} \vec{\beta} - \vec{y} + \gamma \vec{\beta}) = 0$$

$$\vec{\beta} = (\bar{G} + \gamma \bar{I}_n)^{-1} \vec{y}$$

~~coefficients of~~

$$\vec{w}_{RIDGE}^T \vec{x} = \sum_{i=1}^{n} \beta_i \langle \vec{x}_i, \vec{x} \rangle$$