Przypomnienie

Metody selekcji:

- wybór najlepszego podzbioru,
- selekcja krokowa do przodu,
- selekcja krokowa wsteczna.

Metody zmniejszania:

- regresja grzbietowa,
- lasso,
- regresja najmniejszego kąta.

Metody wykorzystujące wyprowadzone kierunki wprowadzania:

- regresja składowych głównych,
- częściowe metody najmniejszych kwadratów.

Least Angle Regression (LAR)

1. Standaryzujemy predyktory tak, aby miały średnią zerową i normę jednostkową. Zaczynamy od reszty $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.
2. Znajdujemy predyktor $\mathbf{x}_j$ najbardziej skorelowany z $\mathbf{r}$.
3. Przesuwamy β_j od 0 w kierunku jego współczynnika najmniejszych kwadratów $\langle \mathbf{x}_j, \mathbf{r} \rangle$, aż inny konkurent $\mathbf{x}_k$ ma taką samą korelację z bieżącą resztą jak $\mathbf{x}_j$.
4. Przesuwamy β_k i β_j w kierunku określonym przez ich łączny współczynnik najmniejszych kwadratów bieżącej reszty na $(\mathbf{x}_j, \mathbf{x}_k)$, aż jakiś inny konkurent $\mathbf{x}_l$ będzie miał taką samą korelację z bieżącą resztą.
5. Kontynuujemy w ten sposób, aż wszystkie predyktory p zostaną wprowadzone. Po krokach min(N − 1,p) dochodzimy do pełnego rozwiązania metodą najmniejszych kwadratów.

$\mathcal{A}_k$ - aktywny zbiór zmiennych na początku k-tego kroku

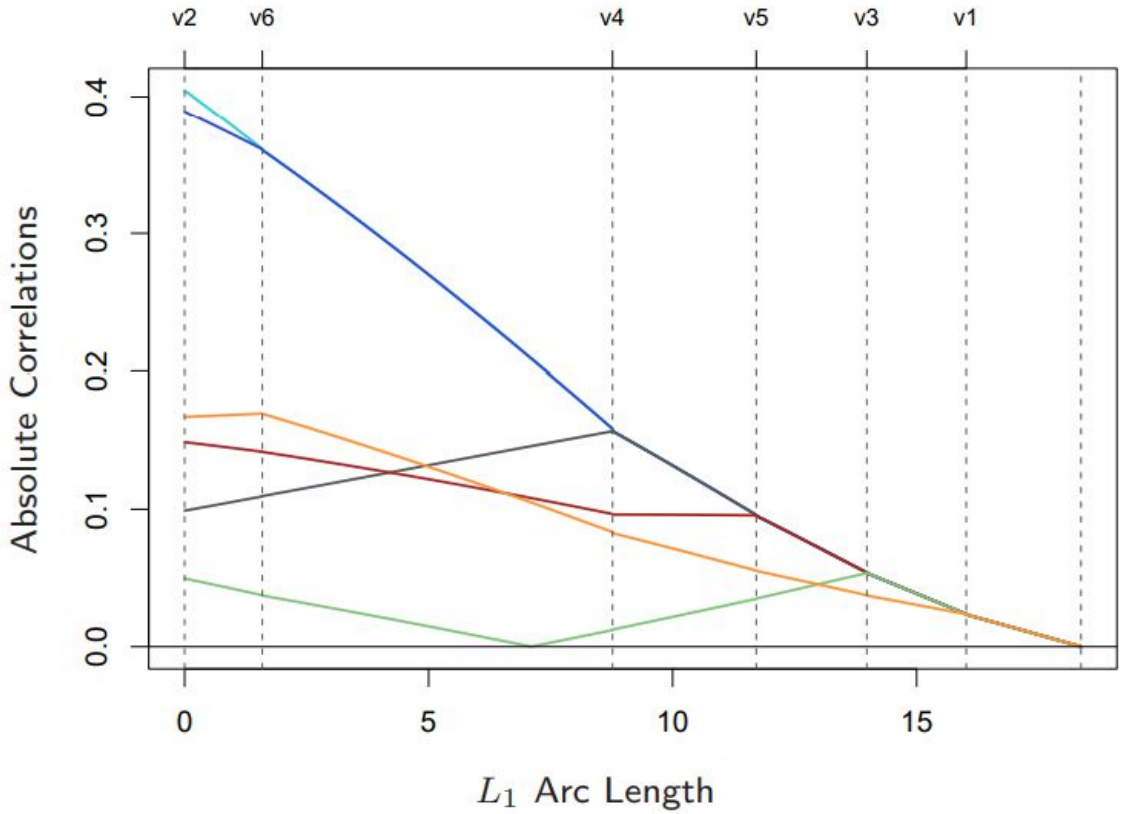
$\beta_{\mathcal{A}_k}$ - wektor współczynników dla tych zmiennych w tym kroku

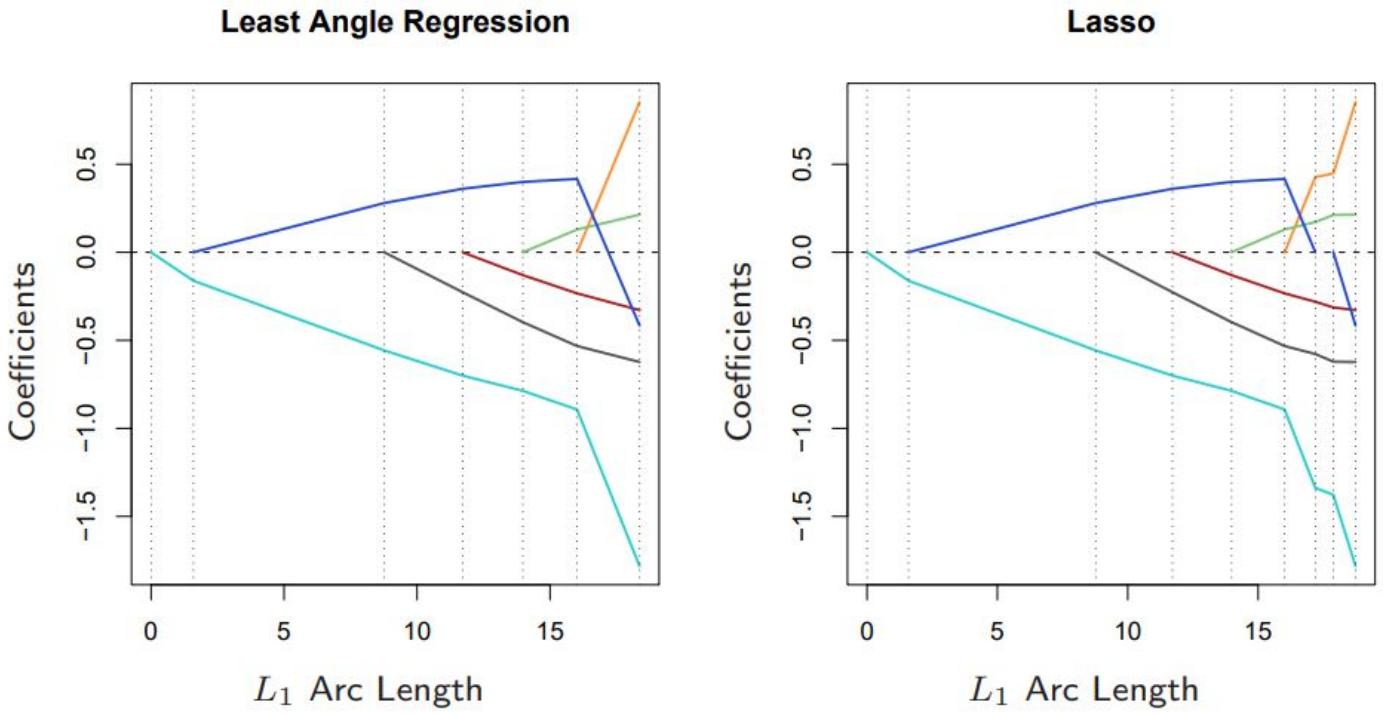
$\mathbf{r}_k = \mathbf{y} - \mathbf{X}_{\mathcal{A}_k}\beta_{\mathcal{A}_k}$ - bieżąca reszta

$\delta_k = (\mathbf{X}_{\mathcal{A}_k}^T \mathbf{X}_{\mathcal{A}_k})^{-1}\mathbf{X}_{\mathcal{A}_k}^T \mathbf{r}_k$ - kierunek dla k-tego kroku

$\beta_{\mathcal{A}_k}(\alpha) = \beta_{\mathcal{A}_k} + \alpha \cdot \delta_k$ - profil współczynnika

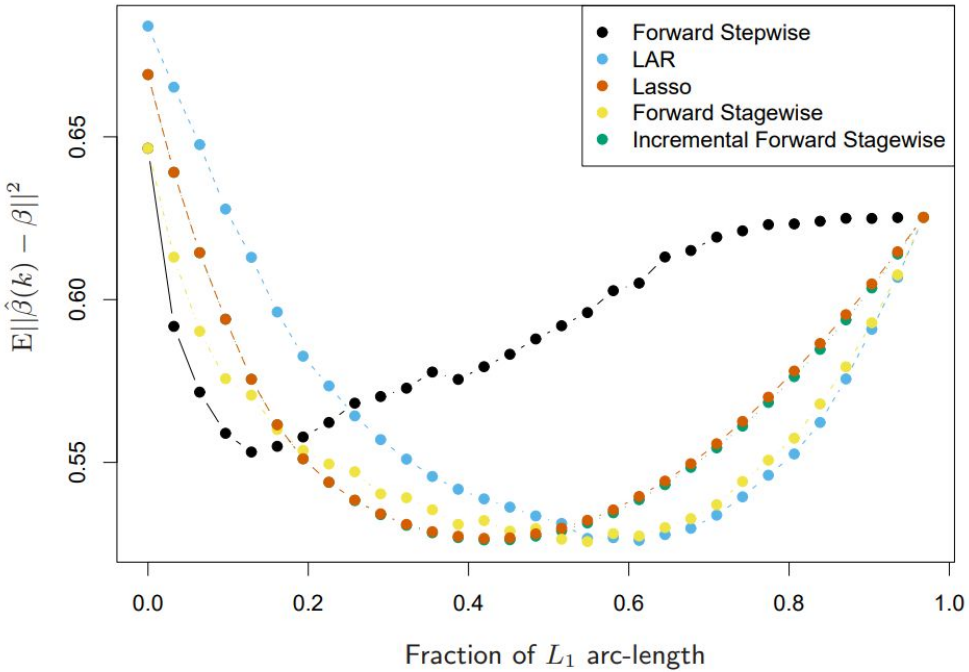
Jeśli wektorem dopasowania na poczatku k-tego kroku jest $\hat{\mathbf{f}}_k$ to ewoluuje jako $\hat{\mathbf{f}}_k(\alpha) = \hat{\mathbf{f}}_k + \alpha \cdot \mathbf{u}_k$, gdzie $\mathbf{u}_k = \mathbf{X}_{\mathcal{A}_k}\delta_k$ jest nowym kierunkiem dopasowania.


v2
v6
v4
v5
v3
v1
Absolute Correlations
0.0
0.1
0.2
0.3
0.4
0
5
10
15
L_1 Arc Length


Least Angle Regression
Coefficients
0.5
0.0
−0.5
−1.0
−1.5
0
5
10
15
L_1 Arc Length
Lasso
Coefficients
0.5
0.0
−0.5
−1.0
−1.5
0
5
10
15
L_1 Arc Length

Modyfikacja algorytmu

4a. Jeśli niezerowy współczynnik osiągnie zero, usuwamy jego zmienną z aktywnego zestawu zmiennych i ponownie obliczamy bieżący wspólny kierunek najmniejszych kwadratów.


Forward Stepwise
LAR
Lasso
Forward Stagewise
Incremental Forward Stagewise
$\mathrm{E}||\hat{\beta}(k) - \beta||^2$
0.55
0.60
0.65
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of L_1 arc-length

N=100

p=31

Degrees-of-Freedom Formula for LAR and Lasso

$$\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$$

$$\mathrm{df}(\hat{\mathbf{y}}) = \frac{1}{\sigma^2} \sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i)$$

Methods Using Derived Input Directions

X_j - dane wejściowe

$Z_m,\ m = 1, \ldots, M$ - kombinacje liniowe danych wejściowych

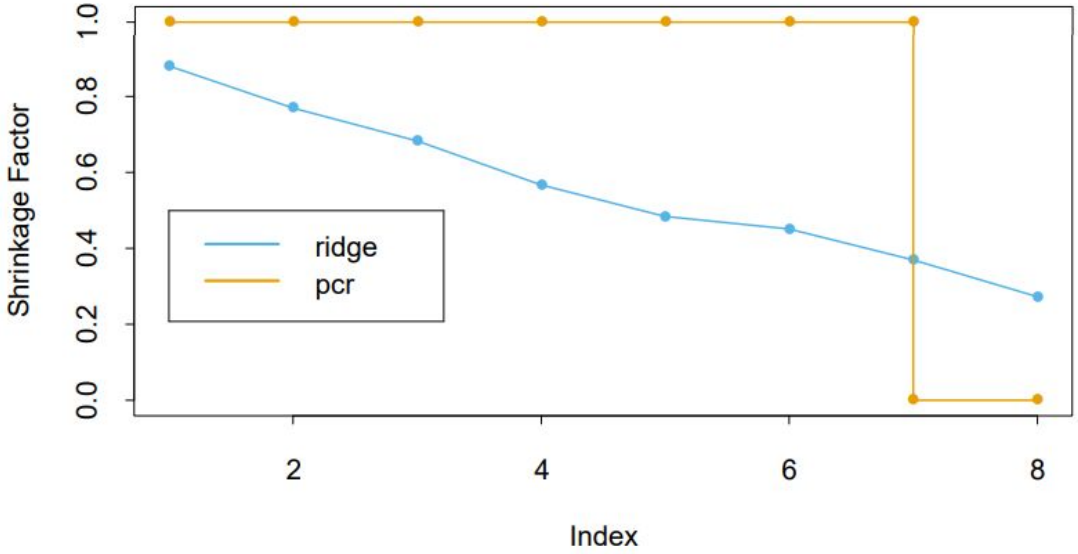
Principal Components Regression (PCR)

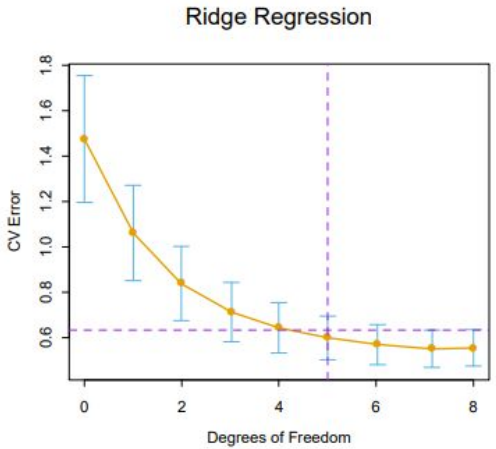
$\mathbf{z}_m = \mathbf{X}v_m$ - pochodne kolumn wejściowych

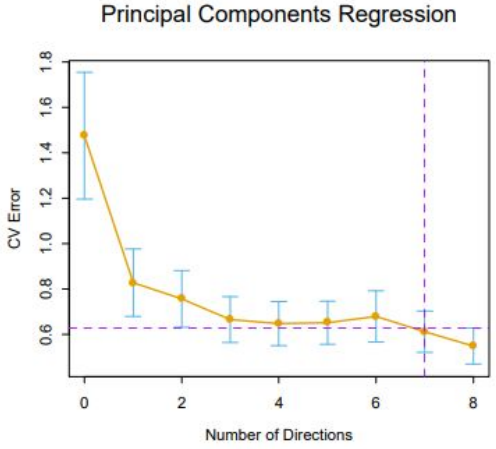
Dokonujemy regresji $\mathbf{y}$ na $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M$ dla pewnego $M \leq p$

$$\hat{\mathbf{y}}^{\text{pcr}}_{(M)} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \mathbf{z}_m \qquad \hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$$

$$\hat{\beta}^{\text{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m$$


Shrinkage Factor
0.0
0.2
0.4
0.6
0.8
1.0
ridge
pcr
2
4
6
8
Index


Ridge Regression
CV Error
0.6
0.8
1.0
1.2
1.4
1.6
1.8
0
2
4
6
8
Degrees of Freedom


Principal Components Regression
CV Error
0.6
0.8
1.0
1.2
1.4
1.6
1.8
0
2
4
6
8
Number of Directions

Partial Least Squares (PLS)

1. Standaryzujemy każdy $\mathbf{x}_j$ tak, aby miał średnią zero i wariancję jeden. Niech $\hat{\mathbf{y}}^{(0)} = \bar{y}\mathbf{1}$ i $\mathbf{x}_j^{(0)} = \mathbf{x}_j,\ j = 1, \ldots, p$.

2. Dla m = 1, 2,... ,p

(a) $\mathbf{z}_m = \sum_{j=1}^{p} \hat{\varphi}_{mj}\mathbf{x}_j^{(m-1)}$, gdzie $\hat{\varphi}_{mj} = \langle \mathbf{x}_j^{(m-1)}, \mathbf{y} \rangle$

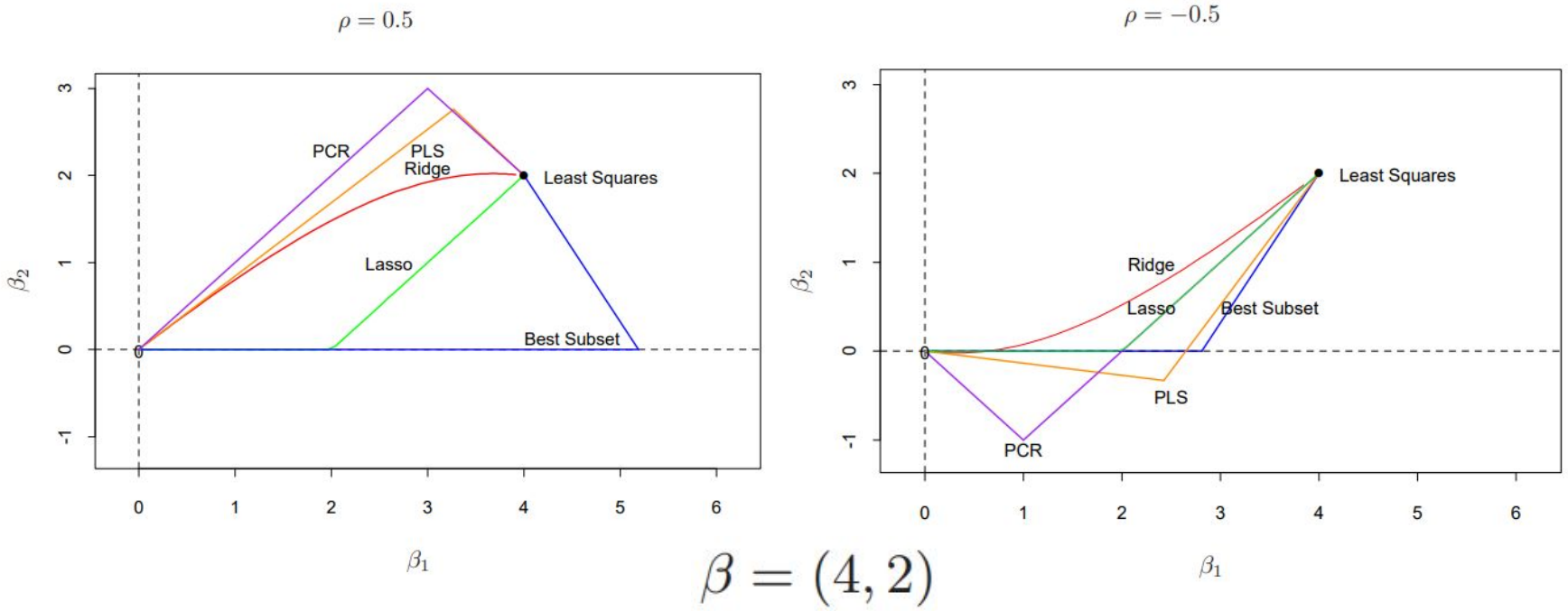
(b) $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$

(c) Ortogonalizujemy każdy $\mathbf{x}_j^{(m-1)}$ względem $\mathbf{z}_m$:

$$\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - [\langle \mathbf{z}_m, \mathbf{x}_j^{(m-1)} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle]\mathbf{z}_m,\ j = 1, 2, \ldots, p.$$

3. Wyprowadzamy sekwencję dopasowanych wektorów $\{\hat{\mathbf{y}}^{(m)}\}_1^p$. Ponieważ $\{\mathbf{z}_\ell\}_1^m$ są liniowe w oryginalnym $\mathbf{x}_j$, więc $\hat{\mathbf{y}}^{(m)} = \mathbf{X}\hat{\beta}^{\text{pls}}(m)$. Te współczynniki liniowe można odzyskać z sekwencji transformacji PLS.

Porównanie metod selekcji i zmniejszania



$$\beta = (4, 2)$$