Shrinkage methods

Metody „kurczeniowe”?

Igor Adamiec

Co to?

- W przeciwieństwie do metod obcinających część zmiennych, shrinkage methods używają wszystkich dostępnych (p) predyktorów,
- Zamiast usuwać zmienne, metody te ograniczają ich współczynniki (coefficients, beta) do 0,
- Skurczenie współczynników może zauważalnie zredukować wariancję,
- Wyróżniamy dwie takie metody: Ridge regressio (regresja grzbietowa?) i lasso regression.

Ridge Regression

Model zwykłej regresji najmniejszych kwadratów szuka takich współczynników by zminimalizować RSS

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

Ridge regression szuka takich współczynników by zminimalizować:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

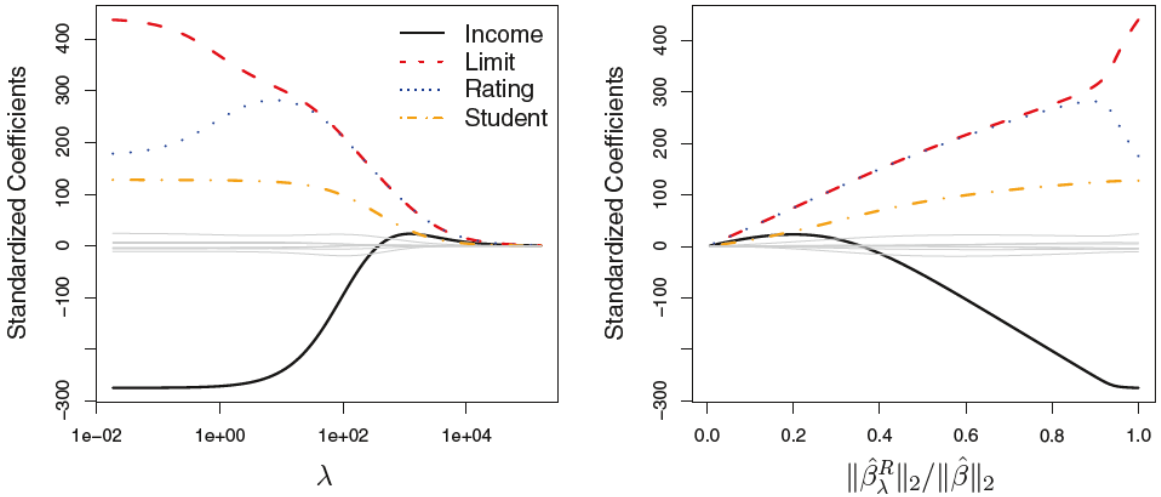
$\lambda \sum_{j=1}^{p} \beta_j^2,$ Shrinkage penalty – mała gdy współczynniki są bliskie zeru

Lambda,
Tuning parameter,
zależy od nas,
>= 0

Suma kwadratów wszystkich współczynników

Gdy lambda jest równa 0, to nie ma żadnej kary.
Im bardziej lambda rośnie, tym kara też rośnie.
Ridge regression tworzy zupełnie inne wartości współczynników dla każdej lambdy.

Ważne: Kara jest tylko dla $\beta_1 \ldots \beta_p$, nie dla β_0



$$\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} {\beta_j}^2}$$

FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

Różne zmienne mogą mieć różne wielkości (np. odległość od miejsca w metrach i liczba posiadanych dzieci – mogą się różnić o kilka rzędów wielkości). Z tego powodu w ridge regression trzeba zestandaryzować (albo znormalizować) dane.

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}},$$

Standaryzacja,
Wartość/ odchylenie

$$\frac{X - \mu}{\sigma}$$

normalizacja

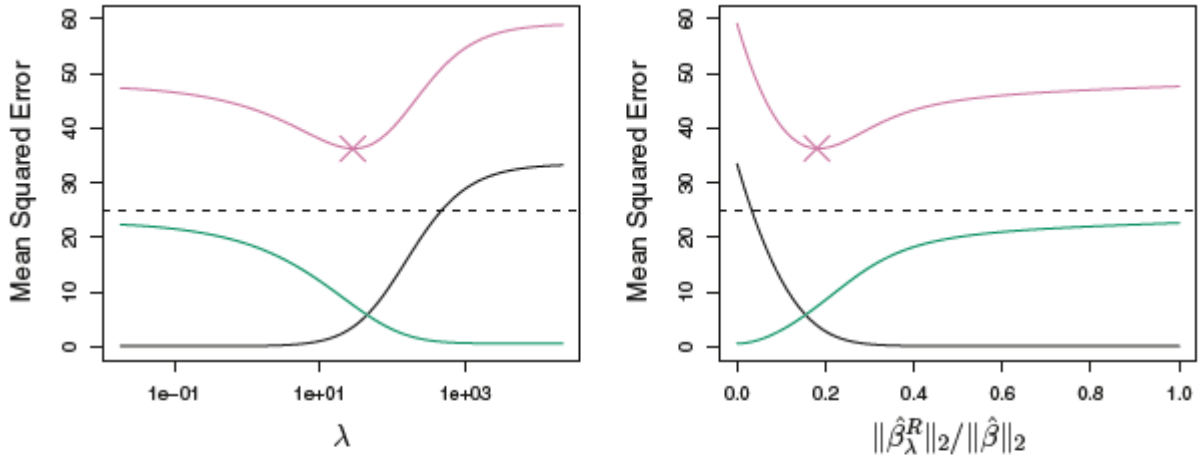


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Lasso regression

- Ridge regression jedynie zmniejszała współczynniki, ale cały czas używała wszystkich zmiennych – bo kara obniżała praktycznie do zera, ale nigdy dokładnie do zera,
- Lasso regresssion pozwala na pozbycie się niektórych zmiennych z modelu,
- W lasso regression szukamy takich współczynników, które zminimalizują:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

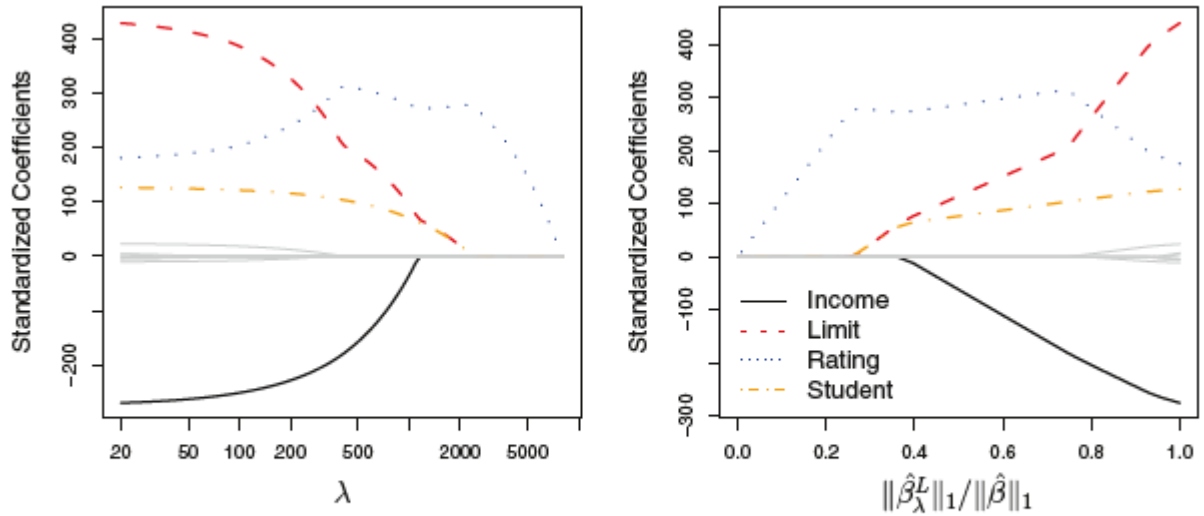


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1$.

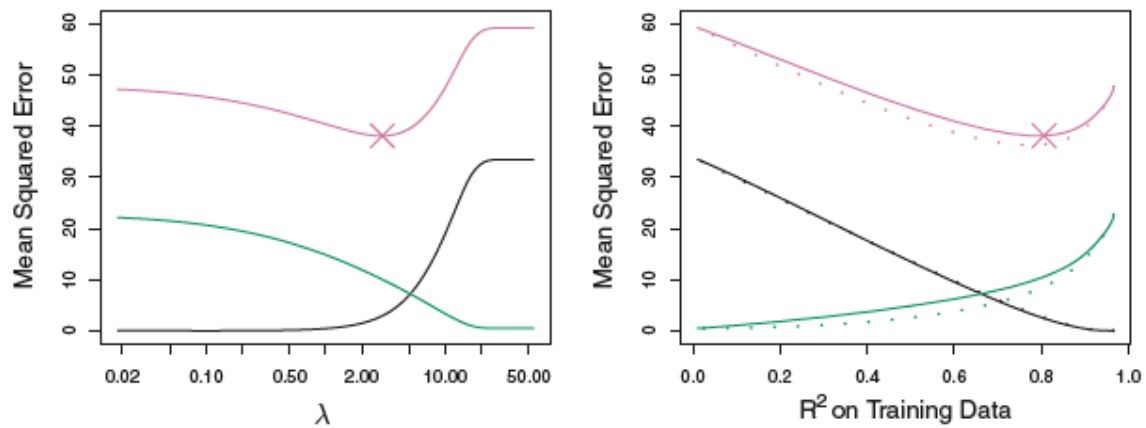


FIGURE 6.8. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their* R^2 *on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

Dlaczego ridge ogranicza, a lasso usuwa zmienne

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s \tag{6.8}$$

lasso

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s, \tag{6.9}$$

ridge

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s. \tag{6.10}$$

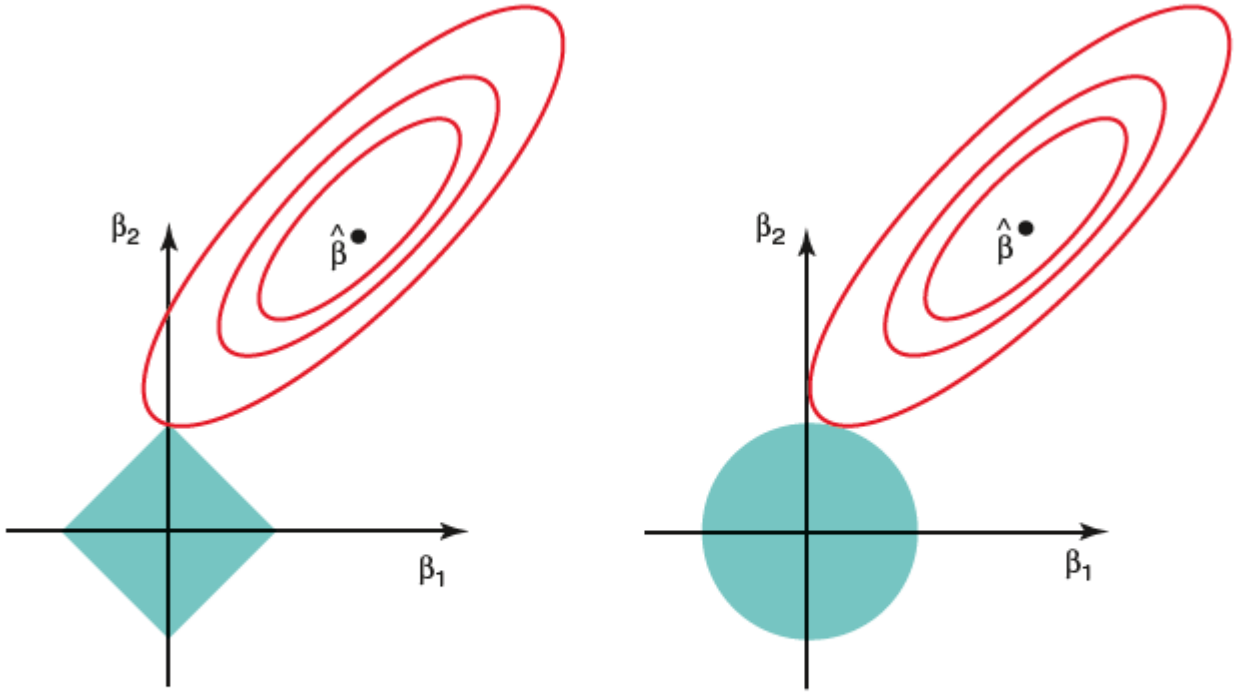


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Dla większej ilości wymiarów, w grę wchodzą inne figury (sfera itd.)

Minimami są punkty styku czerwonej elipsy z zieloną figurą.
Jako, że lasso ma kształt diamentu, istnieje duża szansa, że punkt styku będzie na osi – zmienna będzie równa 0.
Dla ridge (jako, że jest to koło), raczej nie ma takiej możliwości

Ridge vs Lasso

- Ridge osiąga lepsze statystyki gdy nasze y zależy od wszystkich zmiennych (3 slajdy wcześniej),
- Lasso jest lepsze gdy tylko kilka spośród wszystkich zmiennych odpowiada za y,
- Tak naprawdę tego nie wiemy i musimy testować.

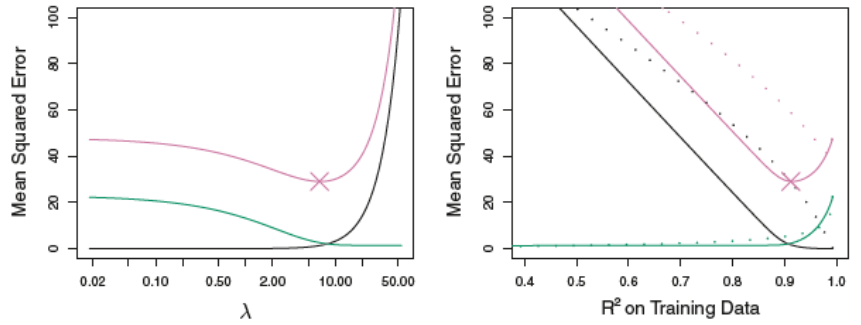


FIGURE 6.9. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Figure 6.8, except that now only two predictors are related to the response.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their R^2 on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

Wybieranie odpowiedniej lambdy

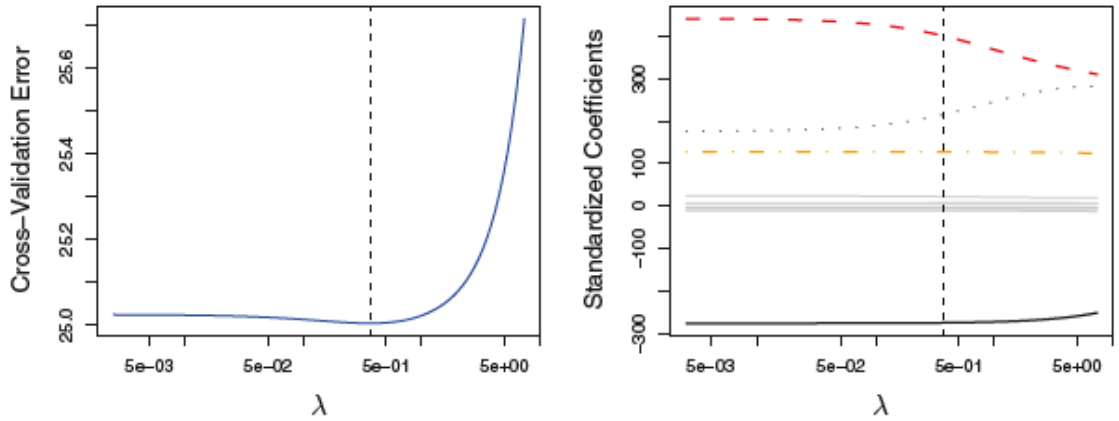


FIGURE 6.12. Left: *Cross-validation errors that result from applying ridge regression to the* `Credit` *data set with various value of* λ. Right: *The coefficient estimates as a function of* λ. *The vertical dashed lines indicate the value of* λ *selected by cross-validation.*

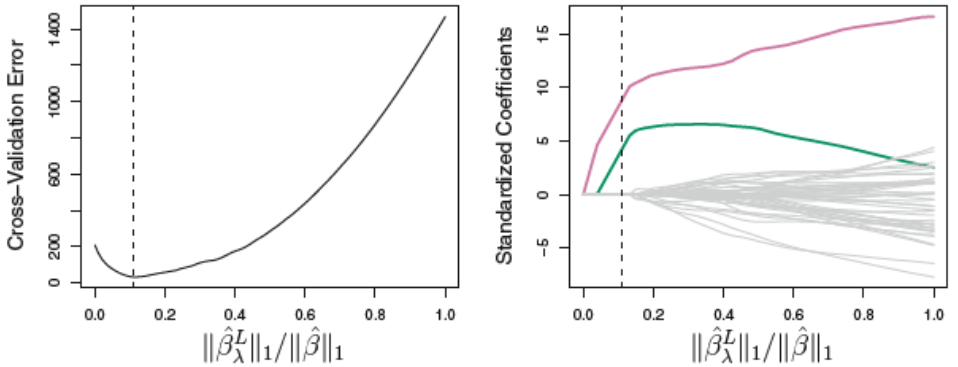


FIGURE 6.13. Left: *Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Figure 6.9.* Right: *The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.*

Metody zmniejszające wymiarowość

Kombinacja liniowa: wektor a) [1,2,3], wektor b) [4, 5,6] kombinacja liniowa a) i b)
1 * 4 + 2* 5 + 3* 6 = 4 + 10 + 18 = 32

Niech $Z_1, Z_2 \dots Z_M, gdzie\ M < p$ reprezentują kombinację liniową naszych p predyktorów.

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \qquad (6.16)$$

$$dla\ stałych\ \phi_{1m}, \phi_{2m} \dots \phi_{pm}, m = 1, \dots, M$$

Możemy wtedy dopasować model metodą najmniejszych kwadratów:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \dots, n,$$

Teraz zamiast przeiwdywać p + 1 współczynników (beta 0, …, beta p), możemy przewidywać tylko M + 1 współczynników

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij},$$

where

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}.$$

Gdy wybrane przez nas M jest dużo mniejsze niż p, to możemy znacznie ograniczyć wariancję. Natomiast jeżeli M = p, to redukcja wymiarowości jest równoznaczna z metodą najmniejszych kwadratów na oryginalnych zmiennych.

Redukcja wymiarowości zawsze ma dwa główne stopnie:
1. Wybór Z_1, Z_2,… Z_M,
2. Dopasowanie modelu z M zmiennymi

Principal Component Regression

- PCR – Regresja głównych składowych opiera się na PCA (analizie głównych składowych),
- Pierwszą składową jest ta zmienna, dla której wartości obserwacji się najbardziej różnią.

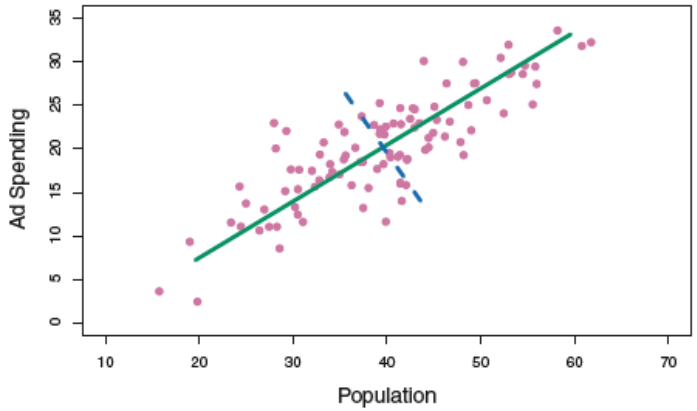


FIGURE 6.14. *The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

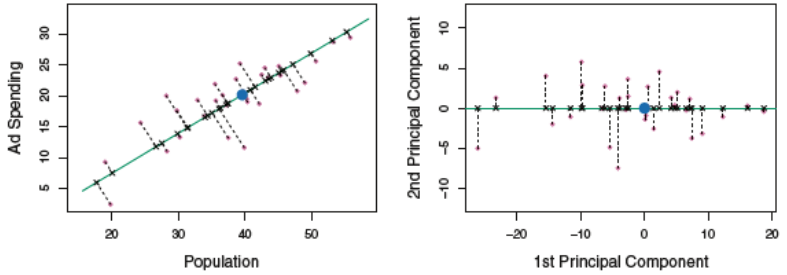


FIGURE 6.15. *A subset of the advertising data. The mean pop and ad budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all n of the observations. The distances from each observation to the principal component are represented using the black dashed line segments. The blue dot represents* $(\overline{\text{pop}}, \overline{\text{ad}})$. Right: *The left-hand panel has been rotated so that the first principal component direction coincides with the x-axis.*

$$Z_1 = 0.839 \times (\text{pop} - \overline{\text{pop}}) + 0.544 \times (\text{ad} - \overline{\text{ad}}).$$

To jest wzór dla pierwszej składowej. Nasze stałe phi równe 0,839 i 0,544 są ładunkami głównych składowych i suma ich kwadratów musi się równać 1. Szukamy takich ładunków, dla których wariancja tego wzoru jest największa.

$$\text{Var}(\phi_{11} \times (\text{pop} - \overline{\text{pop}}) + \phi_{21} \times (\text{ad} - \overline{\text{ad}}))$$

$$z_{i1} = 0.839 \times (\text{pop}_i - \overline{\text{pop}}) + 0.544 \times (\text{ad}_i - \overline{\text{ad}}).$$

Inną definicją PCA jest to, że wektor głównych składowych określa linię, która jest jak najbliższa naszym danym.

Lewy wykres z poprzedniego slajdu sugeruje, że pop i ad są liniowo ze sobą skorelowane. Dlatego też do wyjaśniania tej zależności wystarcza jedna składowa

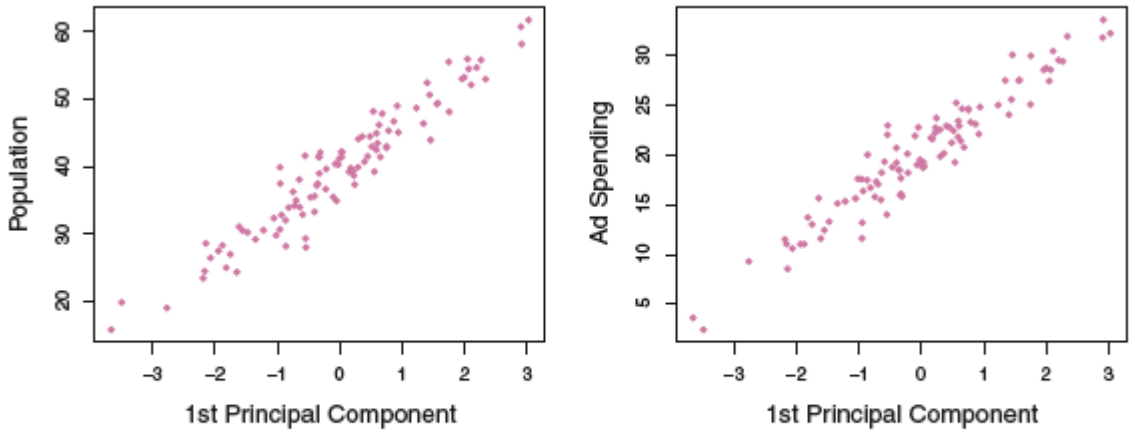


FIGURE 6.16. *Plots of the first principal component scores* z_{i1} *versus* pop *and* ad. *The relationships are strong.*

Druga składowa, Z2, jest kombinacją liniową zmiennych, które są nieskorelowane z Z1 i mają największą wariancję.

Brak korelacji Z2 i Z1 oznacza, że ich kierunki są do siebie prostopadłe.

$$Z_2 = 0.544 \times (\texttt{pop} - \overline{\texttt{pop}}) - 0.839 \times (\texttt{ad} - \overline{\texttt{ad}}).$$

Z2 wykazuje bardzo małą korelację z pop i ad co sugeruje, że wystarczy tylko Z1.

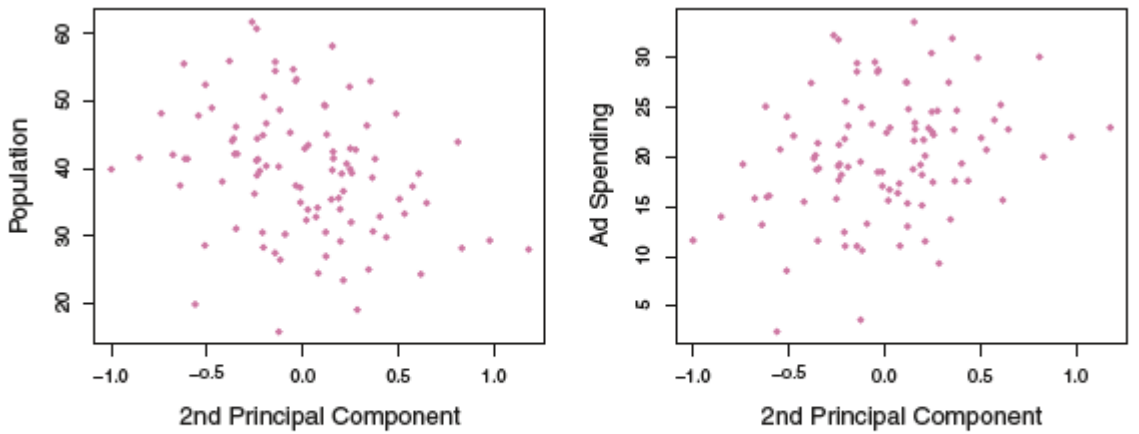


FIGURE 6.17. *Plots of the second principal component scores z_{i2} versus* pop *and* ad*. The relationships are weak.*

PCR zakłada stworzenie M składowych i przy ich użyciu stworzenie modelu regresji najmniejszych kwadratów

Zakładamy, że składowe, dla których zmienne X1,… Xp, wykazują największą wariancję są kierunkami, które są powiązane z Y.

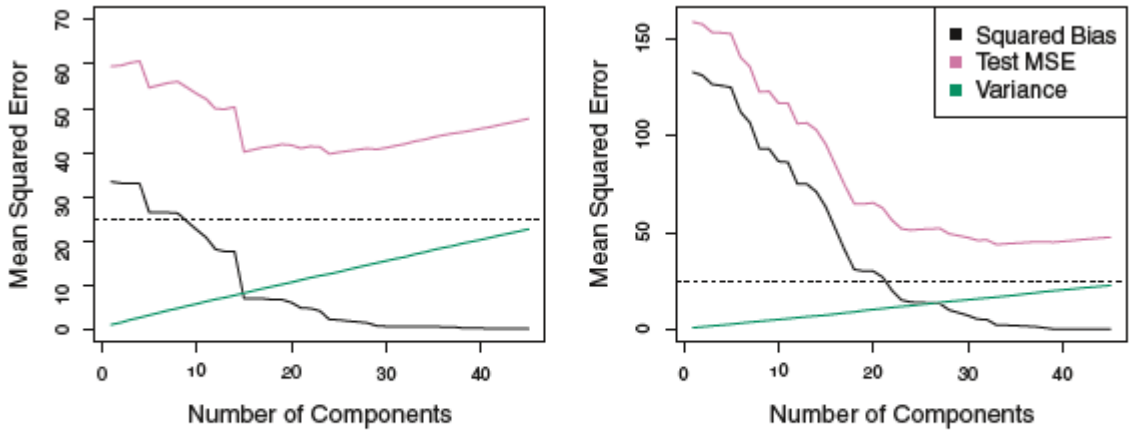


FIGURE 6.18. *PCR was applied to two simulated data sets.* Left: *Simulated data from Figure 6.8.* Right: *Simulated data from Figure 6.9.*

Wszystkie 50 zmiennych jest związane z Y

Tylko 2 z 50 zmiennych są związane z Y

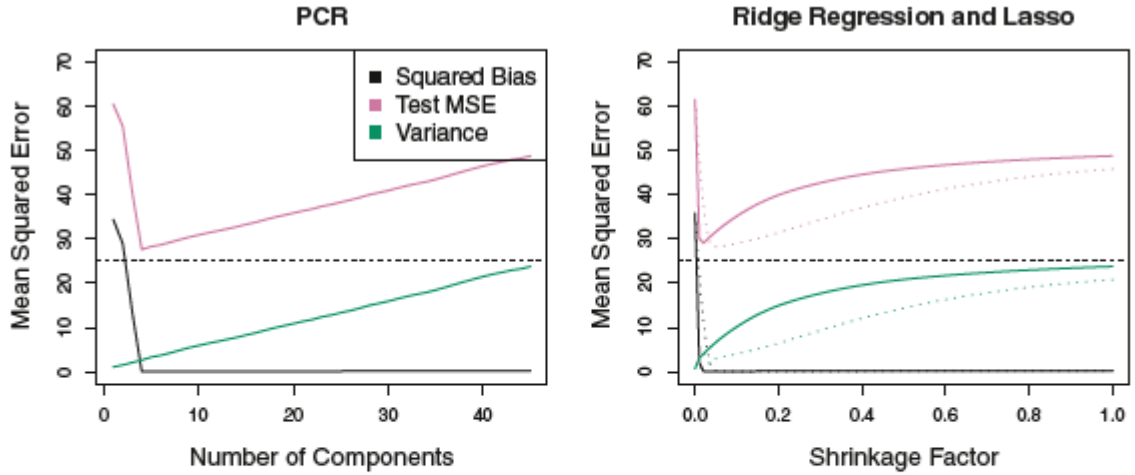


FIGURE 6.19. *PCR, ridge regression, and the lasso were applied to a simulated data set in which the first five principal components of X contain all the information about the response Y. In each panel, the irreducible error $Var(\epsilon)$ is shown as a horizontal dashed line.* Left: *Results for PCR.* Right: *Results for lasso (solid) and ridge regression (dotted). The x-axis displays the shrinkage factor of the coefficient estimates, defined as the ℓ_2 norm of the shrunken coefficient estimates divided by the ℓ_2 norm of the least squares estimate.*

M szukamy przy użyciu CV

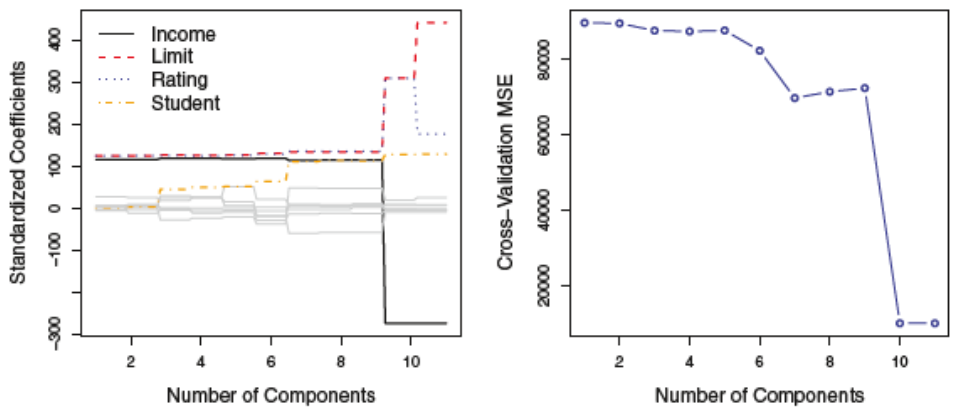


FIGURE 6.20. Left: *PCR standardized coefficient estimates on the* `Credit` *data set for different values of M.* Right: *The ten-fold cross validation MSE obtained using PCR, as a function of M.*

Partial Least Squares

- W PCA składowe są wyliczane bez nadzoru i wartości Y nie są brane pod uwagę przy ich wyliczaniu,
- Z tego powodu nie ma gwarancji, że te wyliczone kierunki najlepiej wyjaśniają zmienną zależną,
- PLS do wyliczania składowych używa również wartości Y.

Wyliczanie Z1 poprzez założenie, że każdy ϕ_1 jest równy współczynnikowi z prostej regresji liniowej Y ~ X_j,

$$Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$$

PLS przykłada największą wagę na zmienne,
Które są najbardziej powiązane z Y.

Na rysunku, większa zmiana jest w zmiennej Pop niż Ad.
To sugeruje, że pop jest bardziej skorelowana z odpowiedzią.

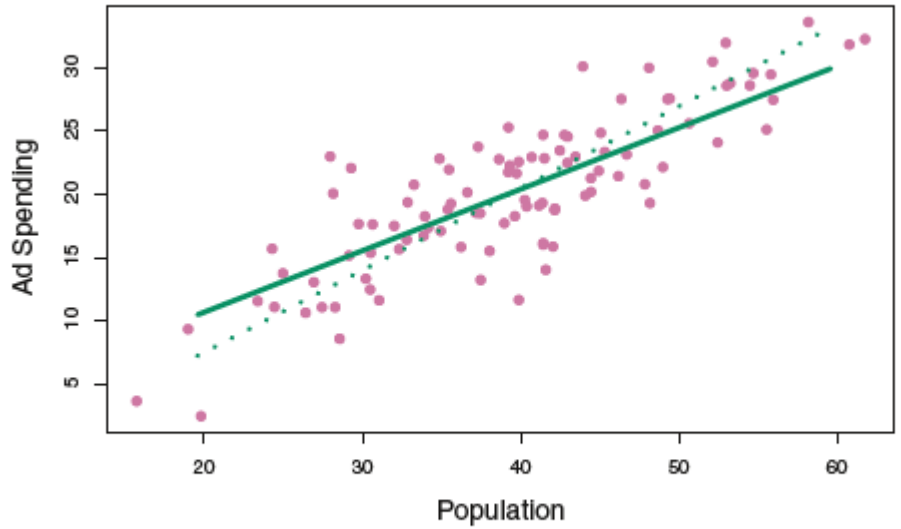


FIGURE 6.21. *For the advertising data, the first PLS direction (solid line) and first PCR direction (dotted line) are shown.*

Drugi kierunek PLS wylicza się poprzez wyliczenie regresji dla Z1 i reszt z tego modelu. Reszty te są interpretowane jako informacje, które nie zostały wyjaśnione przez Z1. Z2 wylicza się na podstawie tych danych na tej samej zasadzie co Z1.