Methods Using Derived Inputs

Statistical Data Mining I

Fall 2017

Rachael Hageman Blair

Shrinkage Methods

- We will tackle a number of shrinkage methods:
  - ~~Forward Stepwise Selection~~
  - ~~Backward Stepwise Selection~~
  - ~~Forward Stagewise Selection~~
  - ~~Ridge Regression~~
  - ~~The LASSO~~
  - ~~Least Angle Regression (LAR)~~
  - Principal Components Regression
  - Partial Least Squares

Selection Methods

Shrinkage Methods

Methods Using Derived Input Directions

Introduction

- **Objective:** derive a reduced set of orthogonal linear projections of a single collection of correlated variables,

$$X = \left(X_1, X_2, \ldots, X_r\right)^T$$

where the projections are ordered by decreasing variances.

- PCA also referred to as a method for decorrelating X, as a result, it has been used in many other fields disguised under different names, e.g, *Karhunrn-Loeve transform (communications theory)* and *empirical orthogonal functions (atmospheric sciences)*.

- Can be used in the supervised and unsupervised setting.

Introduction

- Aside from “dimension reduction” and “de-correlating”, PCA can be used for discovery.

- Discovery takes the form of graphical displays of the principal component scores. The **first few principal component** scores can reveal whether most of the data actually live on a linear subspace, and can be used to:
  - flag outliers
  - discovery anomalies in the distribution
  - identify groupings/ clusters
  - distinguish coordination with experimental treatment factors
- The **last few principal components** also useful from the point of view of outlier detection.

Motivating Example

The Nutritional Value of Food

- 961 food items.
- Nutritional components of each food item are given by the following seven variables: fat (grams), food energy (calories), carbohydrates (grams), protein (grams), cholesterol (milligrams), weight (grams) and saturated fat (grams).
- Food items are listed according to serving sizes, which vary, cup, loaf, piece etc.
- To standardize the different types of servings, the data is scaled and normalized.

food {MMST}

MMST FOOD DATA

Description

nutritional value of food, 196, 198, 206, 208, 462, 612, 613, 631

Usage

```
data(food)
```

Format

A data frame with 961 observations on the following 7 variables.

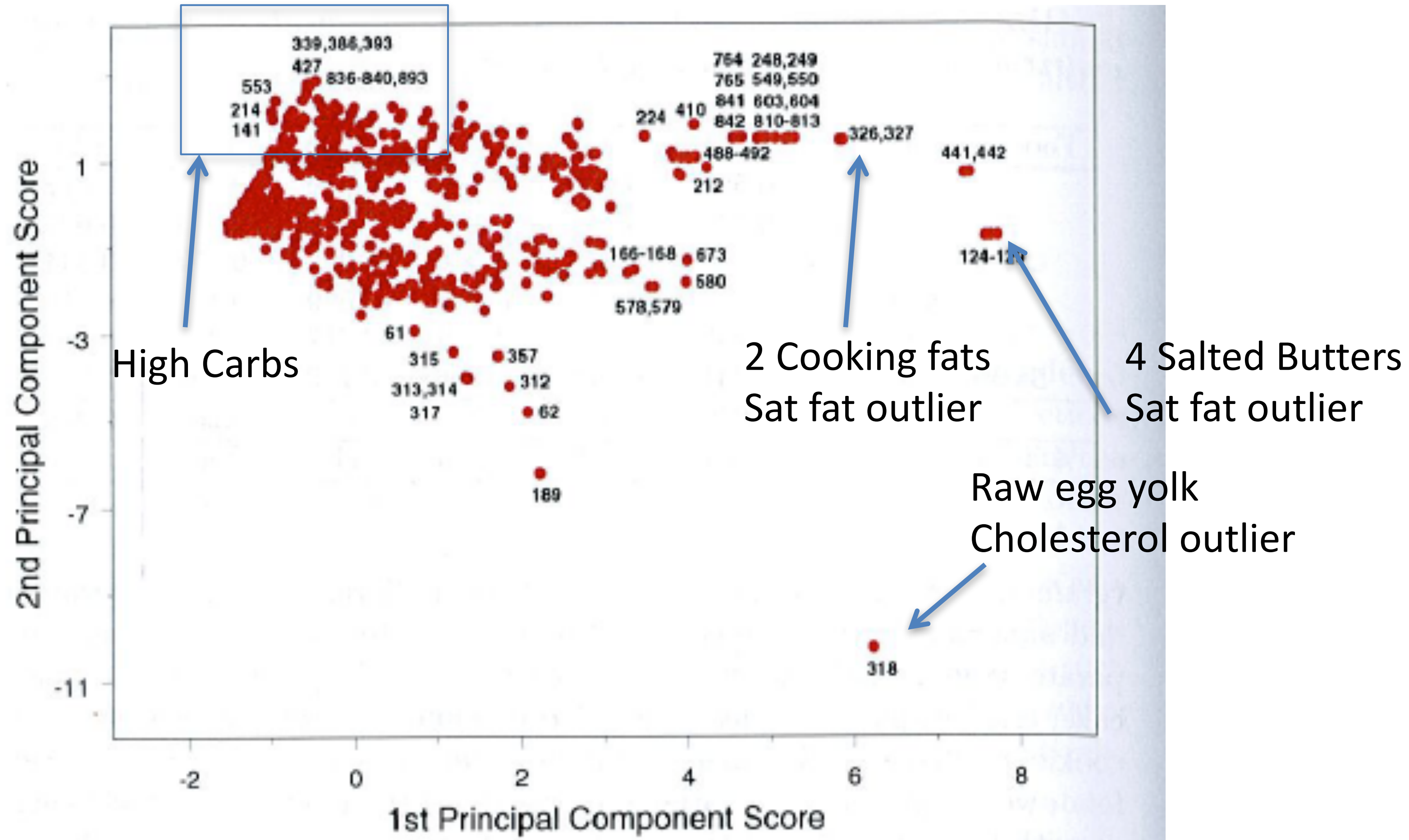
Motivating Example

Before: large matrix 961 x 6

After:

Food Component	PC1	PC2	PC3	PC4	PC5	PC6
Fat	0.557	0.099	0.275	0.130	0.455	0.617
Food energy	0.536	0.357	–0.137	0.075	0.273	-0.697
Carbohydrates	–0.025	0.672	–0.568	–0.286	–0.157	0.344
Protein	0.235	–0.374	–0.639	0.599	–0.154	0.119
Cholesterol	0.253	–0.521	–0.326	–0.717	0.210	–0.003
Saturated fat	0.531	–0.019	0.261	–0.150	-0.791	0.022
Variance	2.649	1.330	1.020	0.680	0.267	0.055
% Total Variance	44.1	22.2	17.0	11.3	4.4	0.9

Motivating Example



Principal Components Regression

General formulation (ISLR):

Let $Z_1, Z_2, \ldots, Z_M$ represent $M < p$ *linear combinations* of our original p predictors. That is,

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \tag{6.16}$$

for some constants $\phi_{1m}, \phi_{2m} \ldots, \phi_{pm}$, $m = 1, \ldots, M$. We can then fit the linear regression model

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n, \tag{6.17}$$

using least squares. Note that in (6.17), the regression coefficients are given by $\theta_0, \theta_1, \ldots, \theta_M$. If the constants $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}$ are chosen wisely, then such dimension reduction approaches can often outperform least squares

Principal Components Regression

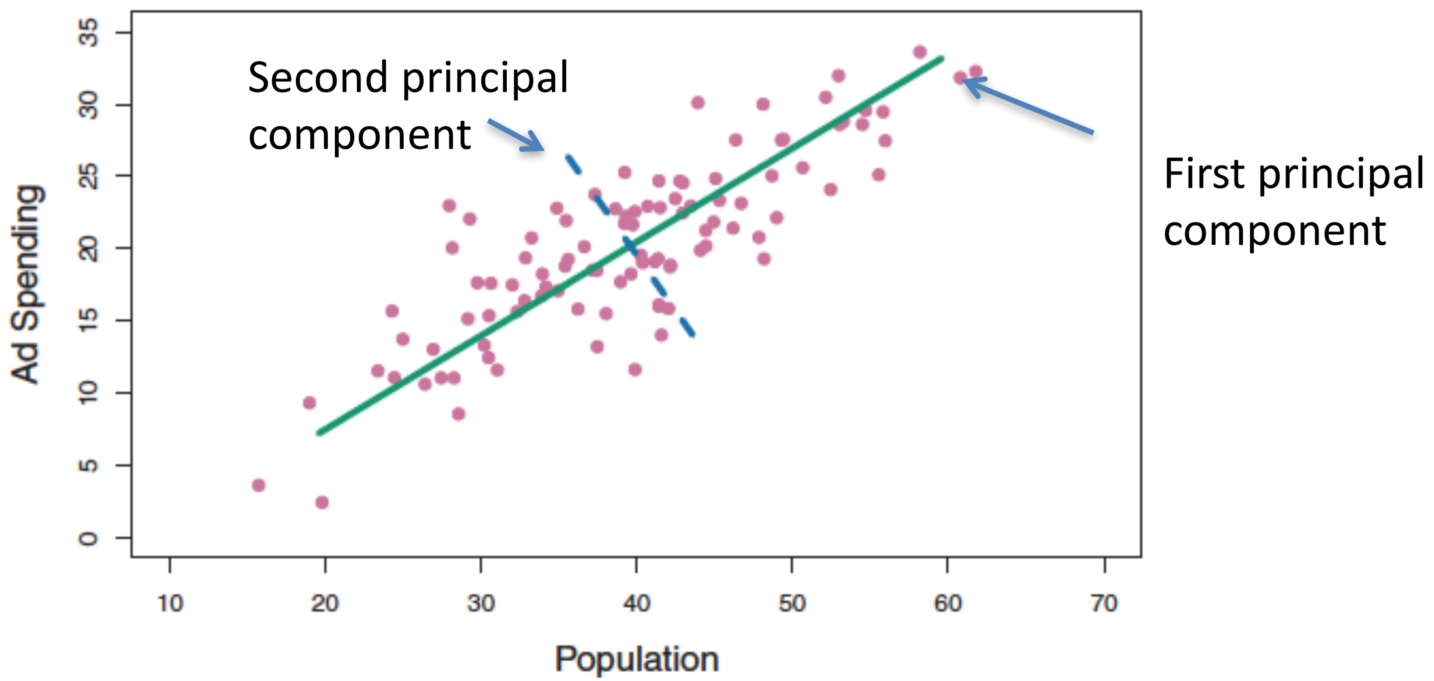


FIGURE 6.14. *The population size (*pop*) and ad spending (*ad*) for* 100 *different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

Principal Components Regression

What is meant by "projecting data"?

- The green line represents the the first principal component direction of the data… greatest variability in the data.
- If we projected 100 observations onto this line, then these would have the most variance (over any competing line).
- The blue line is the second principal component direction, which contains the most information in the data, subject to the fact it has to be orthogonal to PC1.

Projecting a point onto a line →

Finding the location on the line which is

Closest to the point!

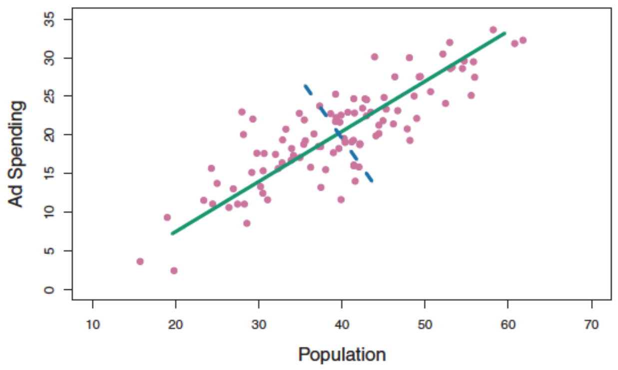


FIGURE 6.14. *The population size* (pop) *and ad spending* (ad) *for* 100 *different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

Principal Components Regression

Mathematically PC1:

$$Z_1 = \underline{0.839} \times \left(pop - \overline{pop} \right) + \underline{0.544} \times \left(ad - \overline{ad} \right)$$

Loadings ϕ_i

Note that

$$\sum_{i=1}^{M} \phi_i = 1$$

*This particular linear combination
Yields the maximum possible variance,
of all linear combinations!

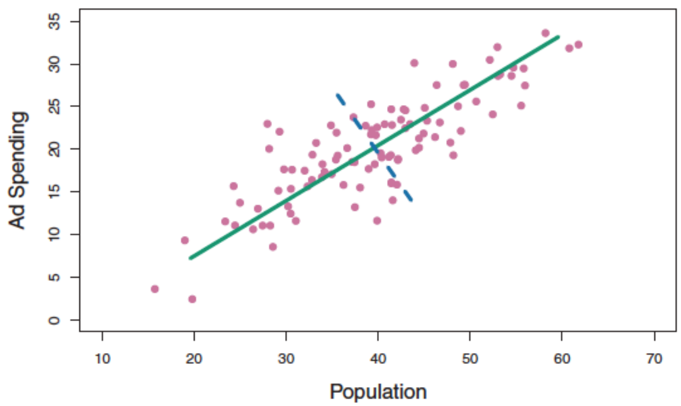


FIGURE 6.14. *The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

Principal Components Regression

Mathematically data points projected to PC1 and PC2:

$$z_{i1} = 0.839 \times \left(pop_i - \overline{pop} \right) + 0.544 \times \left(ad_i - \overline{ad} \right)$$

$$z_{i2} = 0.544 \times \left(pop_i - \overline{pop} \right) - 0.839 \times \left(ad_i - \overline{ad} \right)$$

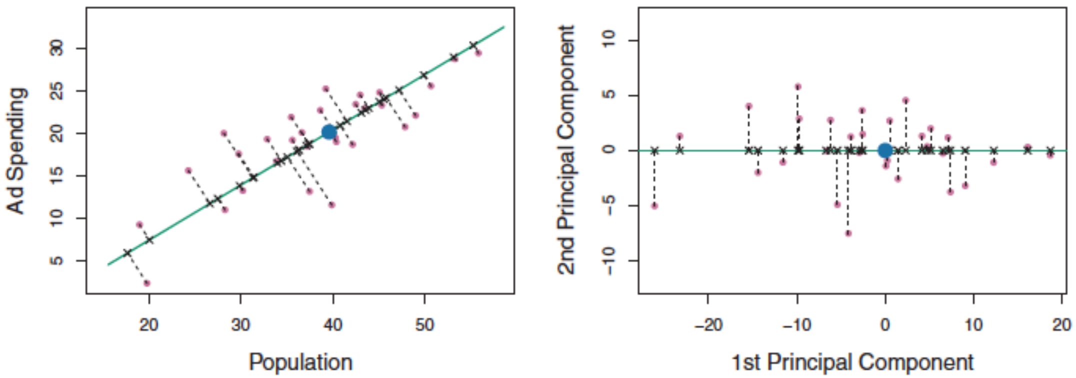


FIGURE 6.15. *A subset of the advertising data. The mean* pop *and* ad *budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all n of the observations. 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.*

Principal Components Regression

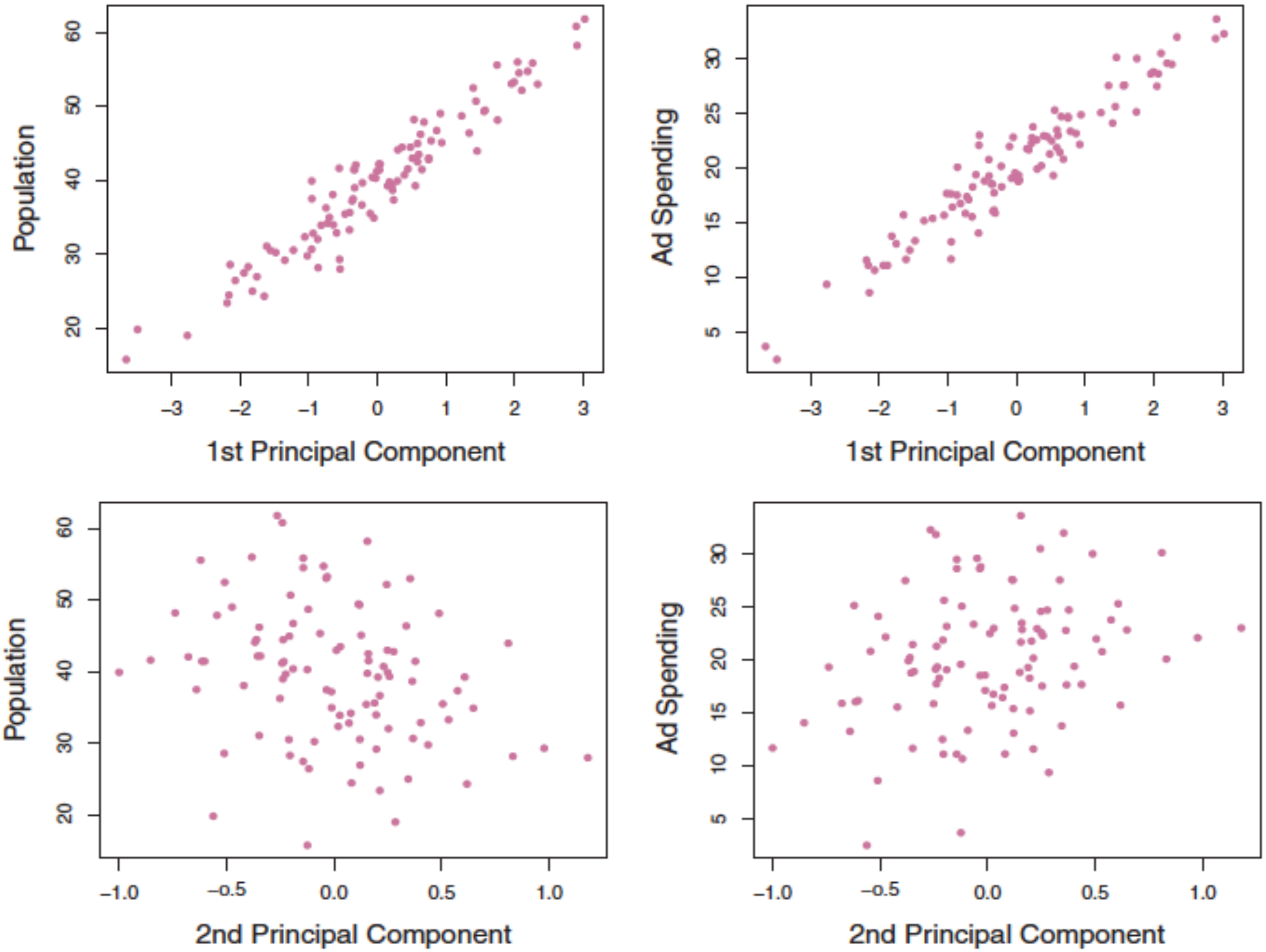
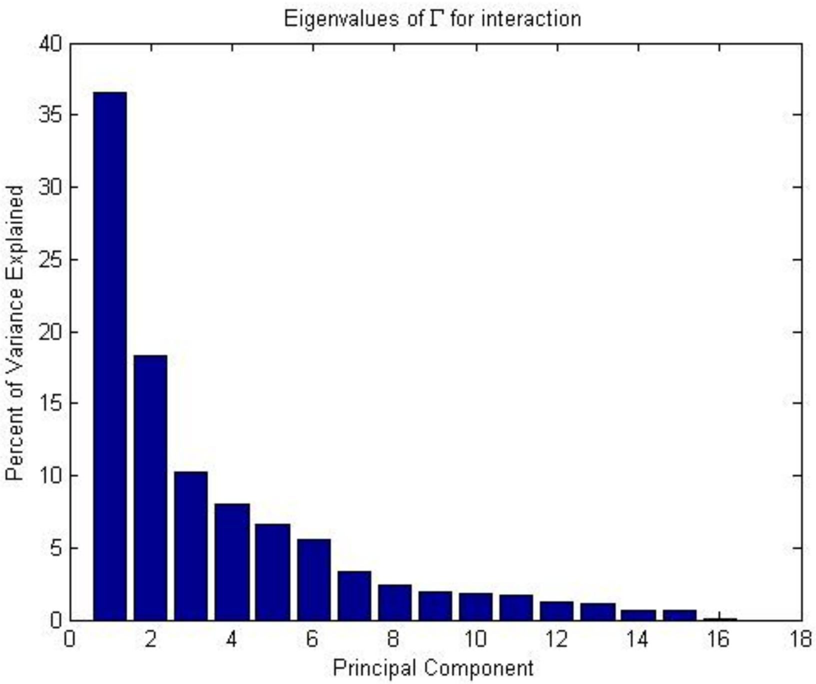
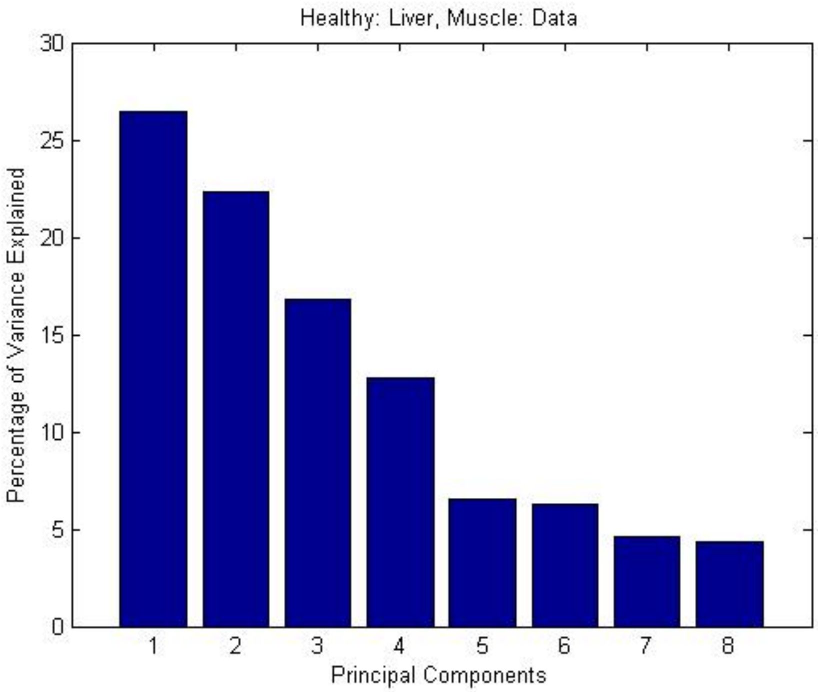


FIGURE 6.17. *Plots of the second principal component scores z_{i2} versus* pop *and* ad*. The relationships are weak.*

Principal Components Regression

Generally, PCA does best when the data can be adequately described by a few components.





Principal Components Regression

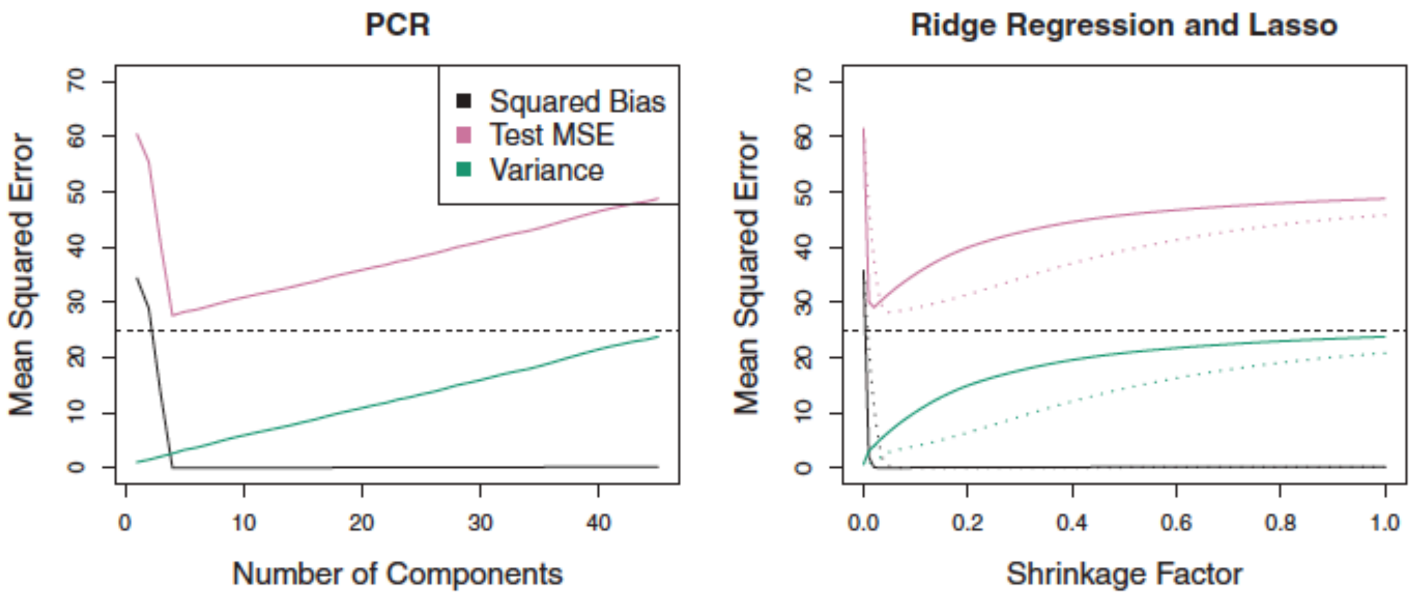


FIGURE 6.19. *PCR, ridge regression, and the lasso were applied to a simulated data set in which the first five principal components of X contain all the information about the response Y. In each panel, the irreducible error $Var(\epsilon)$ is shown as a horizontal dashed line.* Left: *Results for PCR.* 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.*

Principal Components Regression

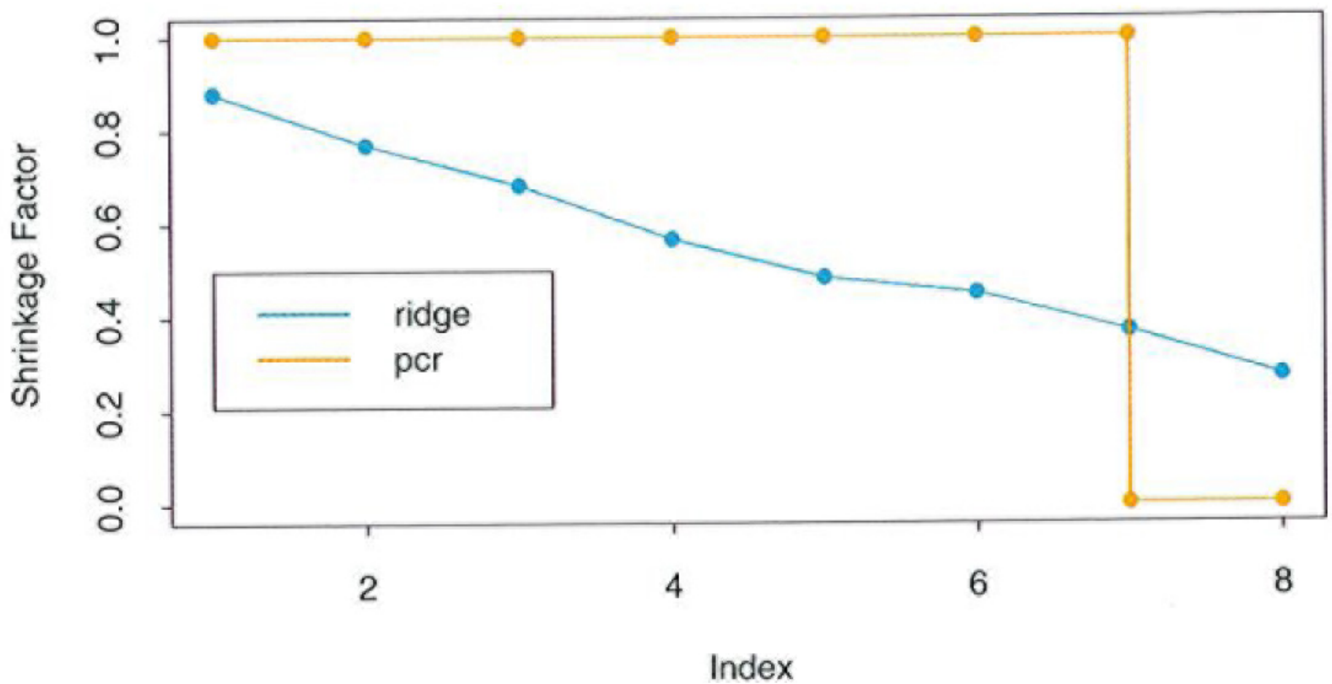


FIGURE 3.17. *Ridge regression shrinks the regression coefficients of the principal components, using shrinkage factors $d_j^2/(d_j^2+\lambda)$ as in (3.47). Principal component regression truncates them. Shown are the shrinkage and truncation patterns corresponding to Figure 3.7, as a function of the principal component index.*

Partial Least Squares Regression

- Note that principal components are chosen to explain X. Nothing guarantees that these chosen components are relevant to Y.

- Partial Least Squares (PLS) seeks directions that have high variance and have high correlation with the response.

- PLS is viewed as being a supervised alternative to PLS.

Partial Least Squares Regression

In a nutshell:

1. The inputs are standardized to have mean 0 and variance 1.

2. The inner product is computed for each *j*: $\hat{\varphi}_{1j} = \langle x_j, y \rangle$.

3. The "derived input" is constructed: $z_1 = \sum_j \hat{\varphi}_{1j} x_j$. This is the first partial least squares direction.

   Therefore, in the construction of each z_m, the inputs are weighted in their univariate effect on y.

Partial Least Squares Regression

In the end:

- Partial least squares produces a sequence of derived, orthogonal inputs or directions: $z_1, z_2, \ldots, z_M$.
- As with PC-regression, if we consider all M=p directions, we are in the case of OLS.

Partial Least Squares- iterative method

Algorithm 3.3 *Partial Least Squares.*

1. Standardize each $\mathbf{x}_j$ to have mean zero and variance one. Set $\hat{\mathbf{y}}^{(0)} = \bar{y}\mathbf{1}$, and $\mathbf{x}_j^{(0)} = \mathbf{x}_j$, $j = 1, \ldots, p$.

2. For $m = 1, 2, \ldots, p$

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

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

Update y hat (c) $\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\theta}_m \mathbf{z}_m$.

Orthogonalize (d) Orthogonalize each $\mathbf{x}_j^{(m-1)}$ with respect to $\mathbf{z}_m$: $\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - [\langle \mathbf{z}_m, \mathbf{x}_j^{(m-1)} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle] \mathbf{z}_m$, $j = 1, 2, \ldots, p$.

Compute the inner product (weights – reflecting the univariate effect on y)

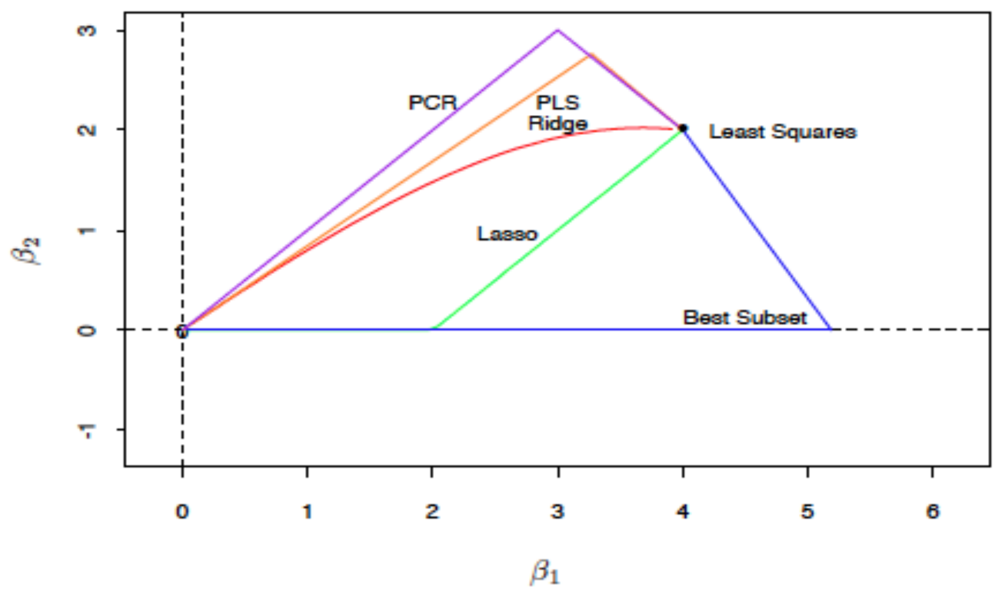
3. Output the sequence of fitted vectors $\{\hat{\mathbf{y}}^{(m)}\}_1^p$. Since the $\{\mathbf{z}_\ell\}_1^m$ are linear in the original $\mathbf{x}_j$, so is $\hat{\mathbf{y}}^{(m)} = \mathbf{X}\hat{\beta}^{\text{pls}}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.

Finer points…

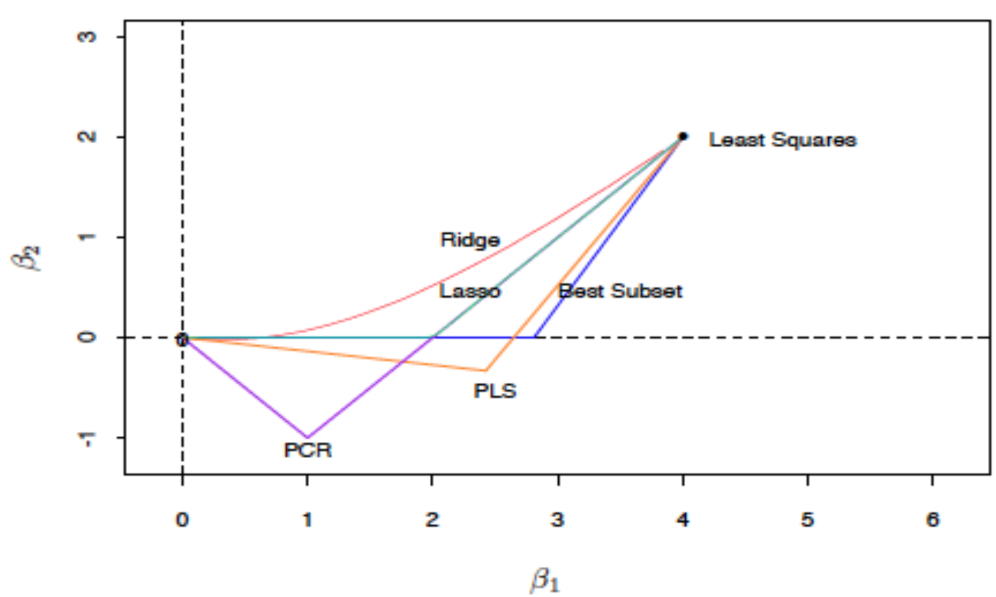
In the end:

- Methods for derived inputs are not feature selection methods.
- Ideas of training and testing have to be thought about carefully.
- PLS in theory, should be superior, but is usually approximate to PCA in practice.

$\rho = 0.5$


PCR
PLS
Ridge
Least Squares
Lasso
Best Subset
β_2
β_1

$\rho = -0.5$


Least Squares
Ridge
Lasso
Best Subset
PLS
PCR
β_2
β_1