DSO530 Statistical Learning Methods

Lecture 6 part III: Dimension Reduction Methods

Dr. Xin Tong
Department of Data Sciences and Operations
Marshall School of Business
University of Southern California

Some Questions

- Q1: What are the “bias” and “variance” trade-off in machine learning?

- Q2: What are the ridge regression and LASSO?

- Q3: Among ridge regression and LASSO, which method tends to give us a parse model?

- Q4: If you have $n = 1,000$ and $p = 2$, is it sensible to apply LASSO? How about ridge regression?

Two ad-hod definitions

- There are many alternative ways to define “bias” and “variance” casually. The follwoing is one of them.
- Bias: the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model
- Variance: the amount by which an estimated model would chance if we were to use a different training set

More About Shrinkage Methods

- We covered shrinkage methods: ridge regression and Lasso
- Note that shrinkage methods are NOT limited these methods
- Other common ones include SCAD, elastic net, etc.
- In the so-called ultra high-dimensional settings (i.e., $p \gg n$), people sometimes use a two-step approach: marginal screening + shrinkage methods

Dimension Reduction Methods

- Subset selection and shrinkage methods use *original* predictors $X_1, \cdots, X_p$
- Dimension reduction methods: pooling the original predictors together to form a few new predictors and then fit a least squares model using the *new* predictors
- Let $Z_1, \cdots, Z_M$ represent $M < p$ linear combinations of our original p predictors. That is

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

  - for some constants $\phi_{1m}, \phi_{2m}, \cdots \phi_{pm}$, and $m = 1, \cdots, M$

- We can then use least squares to fit the linear regression model

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i, \; i = 1, \cdots, n$$

- Usually, $M < p$, so it is called dimension reduction methods
- How to come up with proper ϕ_{jm} is the key

- Notice that

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

- Hence, dimension reduction serves to constrain the estimated β_j coefficients, having the potential to bias the coefficient estimates
- However, what is the upside?
- All dimension reduction methods work in two steps
  - 1. obtain transformed predictors $Z_1, \cdots, Z_M$
  - 2. fit the model using these M predictors
- The ISLR book considers two approaches:
  - principal components
  - partial least squares (not used often)
- We only cover the first one

Principal Components Regression

- Principal components analysis (PCA): a technique for deriving a low-dimensional set of features from a large set of variables
- *first principal component*: the direction of data that along which the observations vary the most

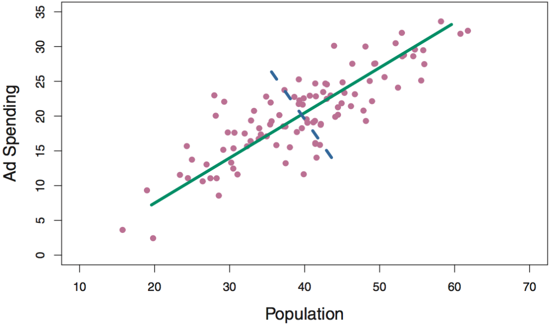

FIGURE 6.14. *The population size (*`pop`*) and ad spending (*`ad`*) for* 100 *different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

- The second principal component Z_2 is a linear combination of the variables that is
  uncorrelated with Z_1, and has largest variance subject to this constraint.

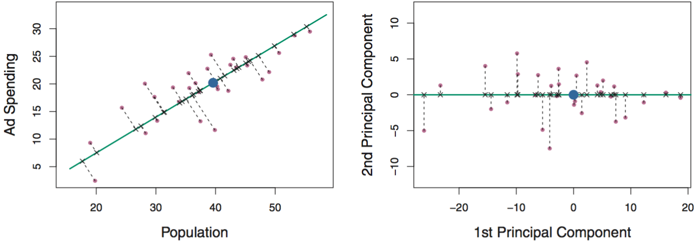

FIGURE 6.15. *A subset of the advertising data. The mean* pop *and* ad *budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all n of the observations.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

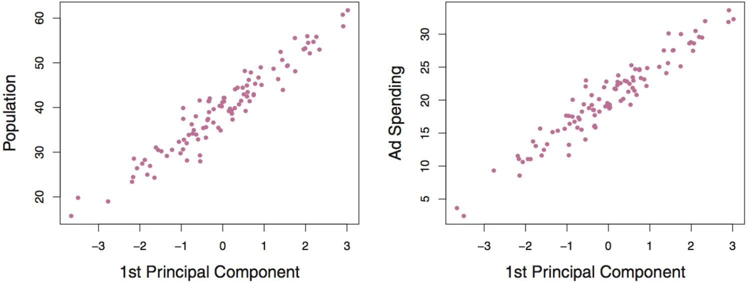

FIGURE 6.16. *Plots of the first principal component scores* z_{i1} *versus* `pop` *and* `ad`*. The relationships are strong.*

Figure 1

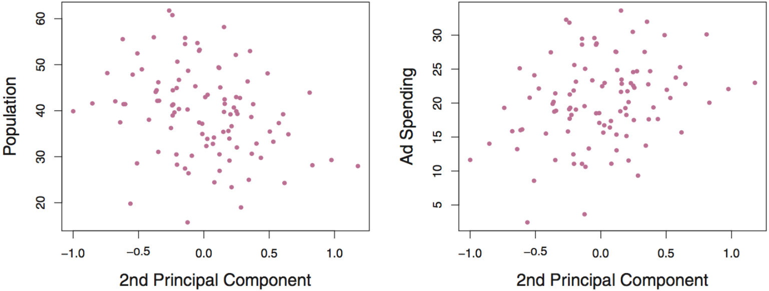

FIGURE 6.17. *Plots of the second principal component scores* z_{i2} *versus* `pop` *and* `ad`*. The relationships are weak.*

Figure 2

Principal Components Regression (PCR)

- Both data have $n = 50$ and $p = 45$

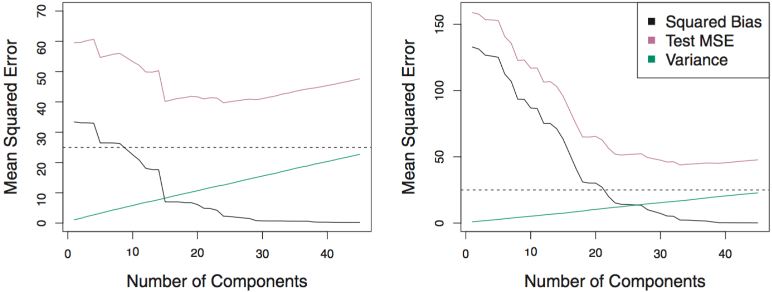

FIGURE 6.18. *PCR was applied to two simulated data sets.* Left: *Simulated data from Figure 6.8.* Right: *Simulated data from Figure 6.9.*

- In PCR, the number of principal components, M, is typically chosen by cross-validation
- When performing PCR, we generally recommend standardizing each predictor, using the formula you saw in part II of Lecture 6
- if the variables are all measured in the same units, then one might choose not to standardize them

Partial Least Squares (PLS) (Optional)

- PCs are obtained in an *unsupervised* way
- That is, the response Y does not *supervise* the identification of the principal components.
- Partial least squares (PLS): a supervised alternative to PCR
- First standardize all p predictors (and the response)
- Then, PLS computes the first direction Z_1 by setting each ϕ_{j1} in equal to the coefficient from the simple linear regression of Y onto X_j. One can show that this coefficient is proportional to the correlation between Y and X_j.
- ϕ_{j1} is proportional to $r(Y, X_j)$

(Optional)

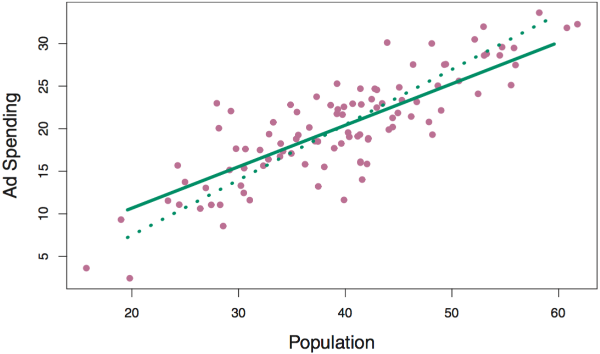

FIGURE 6.21. *For the advertising data, the first PLS direction (solid line) and first PCR direction (dotted line) are shown.*

Figure 3

2nd PLS Direction and Beyond (Optional)

- To identify the second PLS direction, regress each variable on Z_1 and take residuals
- These residuals can be interpreted as the remaining information that has not been explained by the first PLS direction
- We then compute Z_2 using this orthogonalized data in exactly the same fashion as Z_1 was computed based on the original data
- This iterative approach can be repeated M times to identify multiple PLS components $Z_1, \cdots, Z_M$
- M can be chosen by cross-validation
- PLS is good in theory, but not used that often