Dimension Reduction Methods

So far: Two methods for controlling variance of least squares estimators — one uses a subset of the original predictors and the other shrinks their coefficients toward zero. Both use the p original predictors $X_1, \ldots, X_p$.

Now: Two methods that use M *linear combinations* of the p predictors where $M < p$ — **dimension reduction methods**.

- *principal components regression*
- *partial least squares*

But both rely on *principal components analysis* — an **unsupervised learning method** that just works with the data on p variables $X_1, \ldots, X_p$ and there is no associated response variable to supervise. So, let's first learn this topic.

Principal Components Analysis (PCA)

Set up: Have p **correlated** variables $X_1, \ldots, X_p$.

Goal: Get a new set of p **uncorrelated** variables $Z_1, \ldots, Z_p$ — the **principal components** — each of which is a linear combination of the X variables. These are in **decreasing** order of "importance" in that

- Z_1 accounts for as much of the variation as possible amongst all linear combinations of the X variables.
- Z_2 accounts for as much as possible of the remaining variation, subject to being uncorrelated with Z_1, and so on.

Data: n independent subjects giving observations on the p variables, stored as a $n \times p$ matrix $\mathbf{X}$. Assume that the variables are **centered**, i.e., the column means of $\mathbf{X}$ are zero. Also, assume that $p < n$. Let the $p \times p$ matrix $\mathbf{S}$ denote the **sample covariance matrix** of the data. By definition, $\mathbf{S}$ is a *positive semidefinite matrix*. Moreover, $\mathbf{S} = \mathbf{X}^T\mathbf{X}/(n-1)$.

Let (λ_k, ϕ_k), $k = 1, \ldots, p$ denote the eigenvalue-eigenvector pairs of $\mathbf{S}$. In other words, $\mathbf{S}\phi_k = \lambda_k \phi_k$. The eigenvalues are *ordered*, i.e.,

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0,$$

and the eigenvectors are *orthonormal*, i.e.,

$$\phi_k^T \phi_k = 1, \quad \phi_k^T \phi_l = 0, \quad k \neq l.$$

These give the **spectral decomposition** of $\mathbf{S}$, i.e.,

$$\mathbf{S} = \mathbf{\Phi \Lambda \Phi}^T,$$

where $\mathbf{\Phi}$ is a $p \times p$ orthogonal matrix whose columns are the eigenvectors and $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal elements are the eigenvalues.

Let $\phi_{1k}, \ldots, \phi_{pk}$ be the elements of the kth eigenvector ϕ_k. For $k = 1, \ldots, p$, the **kth principal component (PC)** is defined as the linear combination of the variables $X = (X_1, \ldots, X_p)^T$:

$$Z_k = \phi_k^T X = \phi_{1k} X_1 + \ldots + \phi_{pk} X_p.$$

The coefficients ϕ_{jk} are called **loadings**. When the kth PC is evaluated on the ith subject, we get the **score**,

$$Z_{ik} = \phi_{1k} X_{i1} + \ldots + \phi_{pk} X_{ip} = (i\text{th row of } \mathbf{X})\phi_k, \ i = 1, \ldots, n.$$

The scores are values of the PCs on the n subjects. They represent **projections** of the original observations onto the directions determined by the eigenvectors. Let $\mathbf{Z}$ be the $n \times p$ matrix of these scores. Then, we have

$$\mathbf{Z} = \mathbf{X}\mathbf{\Phi},$$

where $\mathbf{\Phi}$ is the matrix of the loadings. Post-multiplying both sides by $\mathbf{\Phi}^T$ and using orthogonality of $\mathbf{\Phi}$ gives,

$$\mathbf{Z}\mathbf{\Phi}^T = \mathbf{X}\mathbf{\Phi}\mathbf{\Phi}^T = \mathbf{X}.$$

Sample mean of Z_k:

$$\sum_{i=1}^{n} Z_{ik} = \sum_{i=1}^{n} (i\text{th row of } \mathbf{X})\phi_k = \left(\sum_{i=1}^{n} (i\text{th row of } \mathbf{X}) \right) \phi_k = 0$$

Sample covariance matrix of $Z_1 \ldots, Z_p$:

$$\frac{\mathbf{Z}^T\mathbf{Z}}{n-1} = \frac{\mathbf{\Phi}^T\mathbf{X}^T\mathbf{X}\mathbf{\Phi}}{n-1} = \mathbf{\Phi}^T\mathbf{S}\mathbf{\Phi} = \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^T\mathbf{\Phi} = \mathbf{\Lambda}$$

- The PC $Z_1, \ldots, Z_p$ are uncorrelated, have mean zero, and their variances are $\lambda_1, \ldots, \lambda_p$ (in decreasing order).
- Total variance of the PC $= \sum_{k=1}^{p} \lambda_k = \text{trace}(\mathbf{S}) =$ total variance of the predictors

Sample covariance between Z_k and X_j: Write $X_j = a_j^T X$ where the jth element of a_j is 1 and the rest are zero. Then,

$$\text{cov}(Z_k, X_j) = \text{cov}(\phi_k^T X, a_j^T X) = \phi_k^T S a_j = \lambda_k \phi_k^T a_j = \lambda_k \phi_{jk}.$$

Optimization Problems Solved by the PC

Let $a^T X$ be a linear combination of the p variables X. The sample mean of $a^T X$ is zero and its sample variance is $a^T S a$. Two linear combinations $a_1^T X$ and $a_2^T X$ are orthogonal if $a_1^T a_2 = 0$. Consider finding the linear combination $a^T X$ that maximizes its variance $a^T S a$. We need to impose a constraint on a otherwise the variance can be made infinitely large. The constraint is $a^T a = 1$.

1st PC: $\phi_1^T X$ is the linear combination that maximizes $a_1^T S a_1$ subject to $a_1^T a_1 = 1$

2nd PC: $\phi_2^T X$ is the linear combination that maximizes $a_2^T S a_2$ subject to $a_2^T a_2 = 1$ and $a_2^T \phi_1 = 0$

***k*th PC**: $\phi_k^T X$ is the linear combination that maximizes $a_k^T S a_k$ subject to $a_k^T a_k = 1$ and $a_k^T \phi_l = 0$ for all $l < k$.

Note: $\text{cov}(a_k^T X, \phi_l^T X) = a_k^T S \phi_l = \lambda_l a_k^T \phi_l$. Therefore, $a_k^T X$ and $\phi_l^T X$ are orthogonal, i.e., $a_k^T \phi_l = 0 \equiv$ they are uncorrelated.

To summarize:

- Since $\mathbf{Z} = \mathbf{X}\mathbf{\Phi}$, the p PC $Z_1, \ldots, Z_p$ are a rotation of the p original variables $X_1, \ldots, X_p$ that preserves the variation (in the sense that their total variances are identical)
- $Z_1, \ldots, Z_p$ are uncorrelated and have variances $\lambda_1, \ldots, \lambda_p$ (in decreasing order).
- The first eigenvector ϕ_1 defines a direction in the p-dimensional feature space along which the data vary the most. Projecting each of the observations onto this direction gives the scores $Z_{11}, \ldots, Z_{n1}$ on the first PC Z_1. The second eigenvector ϕ_2 defines a direction that is perpendicular to the direction ϕ_1 and along which the data vary the most. Projecting each of the observations onto this direction gives the scores $Z_{12}, \ldots, Z_{n2}$ on the second PC Z_2. And so on.

- The first M PC provide the **best** M-dimensional plane that approximates the original observations, i.e.,

$$X_{ij} \approx \sum_{k=1}^{M} Z_{ik}\phi_{jk}.$$

  This approximation is exact when $M = p$. To understand this, recall that $\mathbf{Z} = \mathbf{X}\mathbf{\Phi}$ and the orthogonality of $\mathbf{\Phi}$ gives $\mathbf{X} = \mathbf{Z}\mathbf{\Phi}^T$. Thus, the (i, j)th element of $\mathbf{X}$ is the product of ith row of $\mathbf{Z}$ and jth column of $\mathbf{\Phi}^T$, which is same as jth row of $\mathbf{\Phi}$. If, instead of using all the p PC, only the first M PC are used, we get the above approximation.

- The p eigenvectors provide a **data-driven basis** to represent the data. This allows writing observations as linear combinations of the eigenvectors with scores as the coefficients.

- When $p = 2$, the first PC defines the line that is as close to as possible to the data — in the sense of minimizing SS of perpendicular distances between each point and the line. Projecting observations on this line gives the scores on the 1st PC, which represents the best 1-D summary of the data.

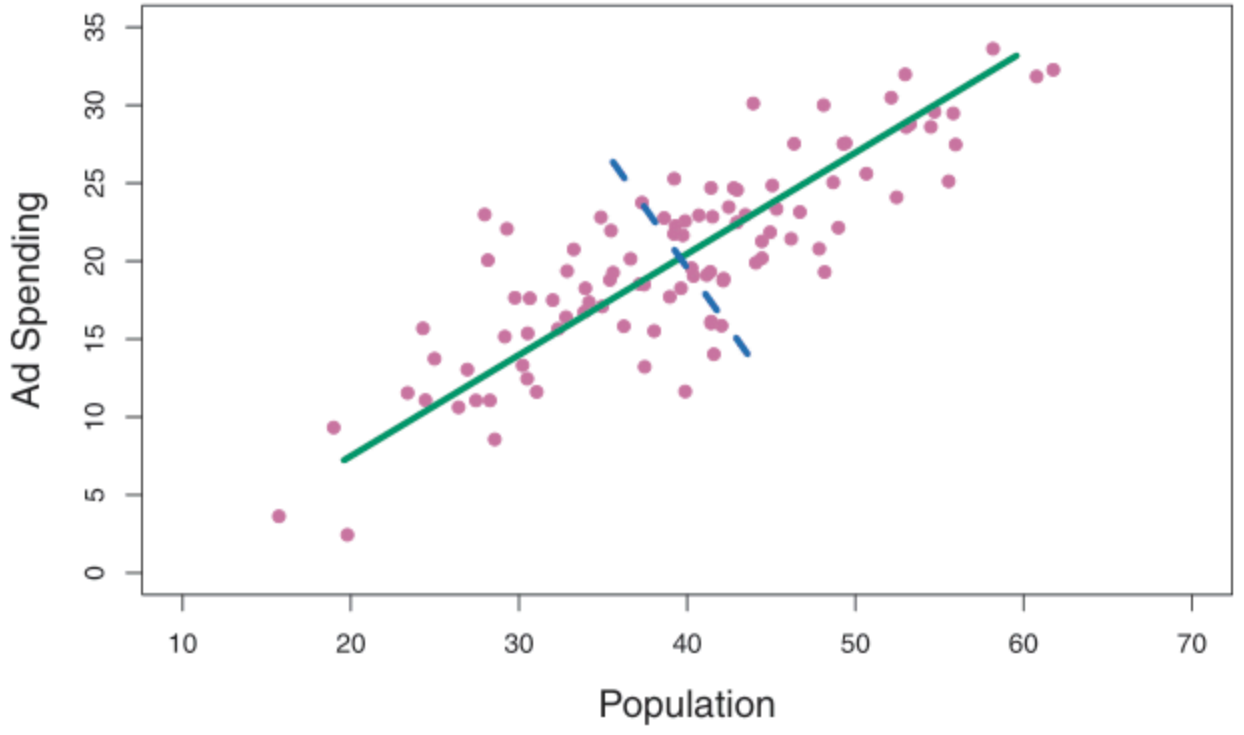

FIGURE 6.14. *The population size (*pop*) and ad spending (*ad*) for* 100 *different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

- $Z_1 = 0.839 \times (\text{pop} - \overline{\text{pop}}) + 0.544 \times (\text{ad} - \overline{\text{ad}})$
- $Z_2 = 0.544 \times (\text{pop} - \overline{\text{pop}}) - 0.839 \times (\text{ad} - \overline{\text{ad}})$

Source: ISL

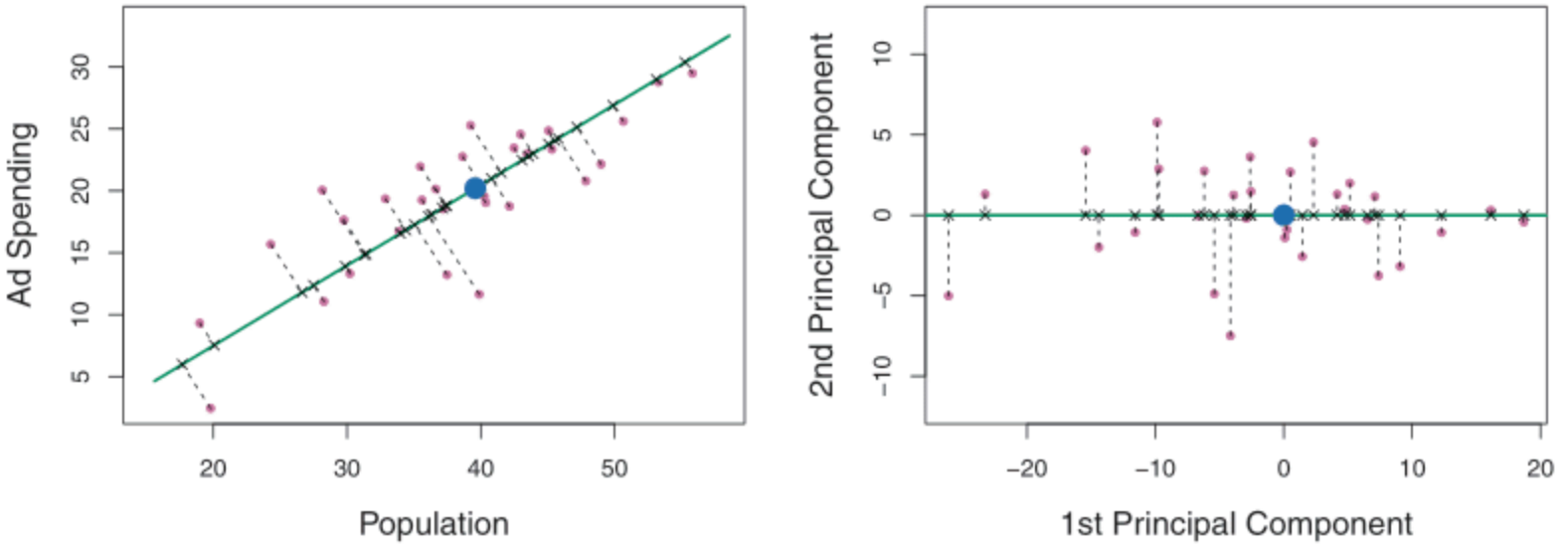

FIGURE 6.15. *A subset of the advertising data. The mean* pop *and* ad *budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all* n *of the observations.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

Source: ISL

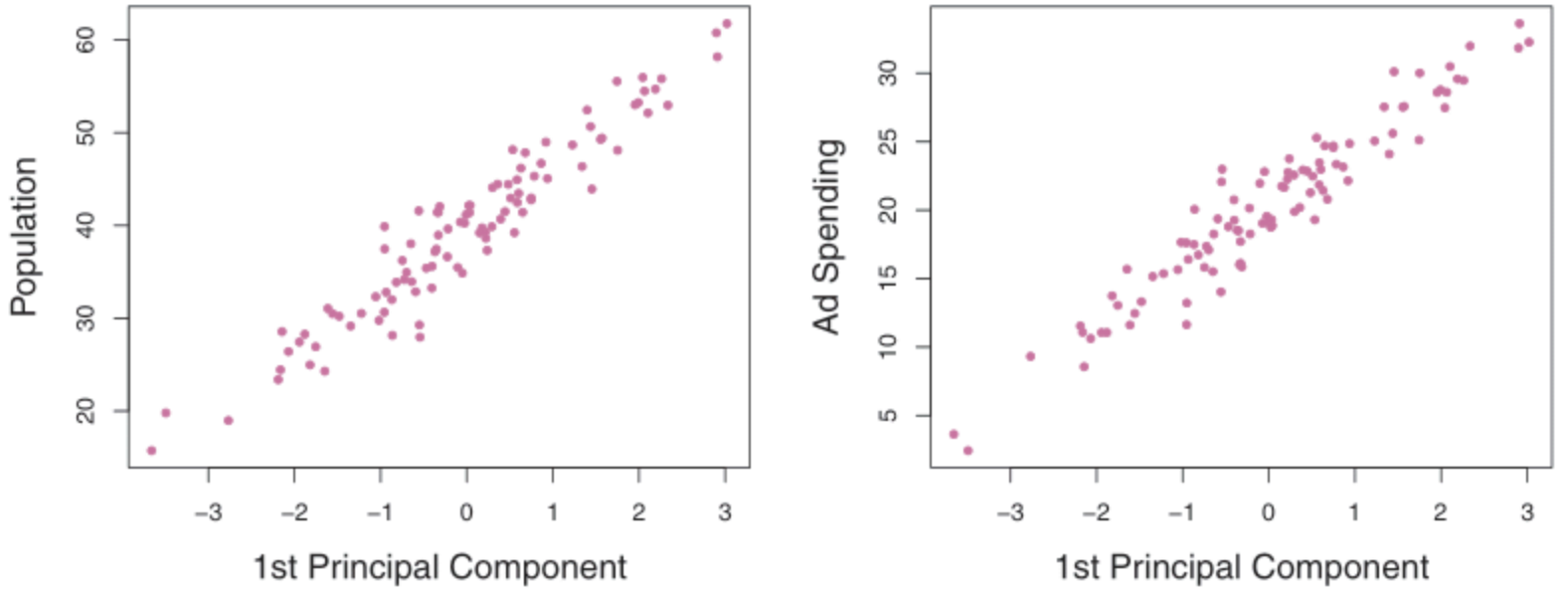

FIGURE 6.16. *Plots of the first principal component scores z_{i1} versus* pop *and* ad*. The relationships are strong.*

Source: ISL

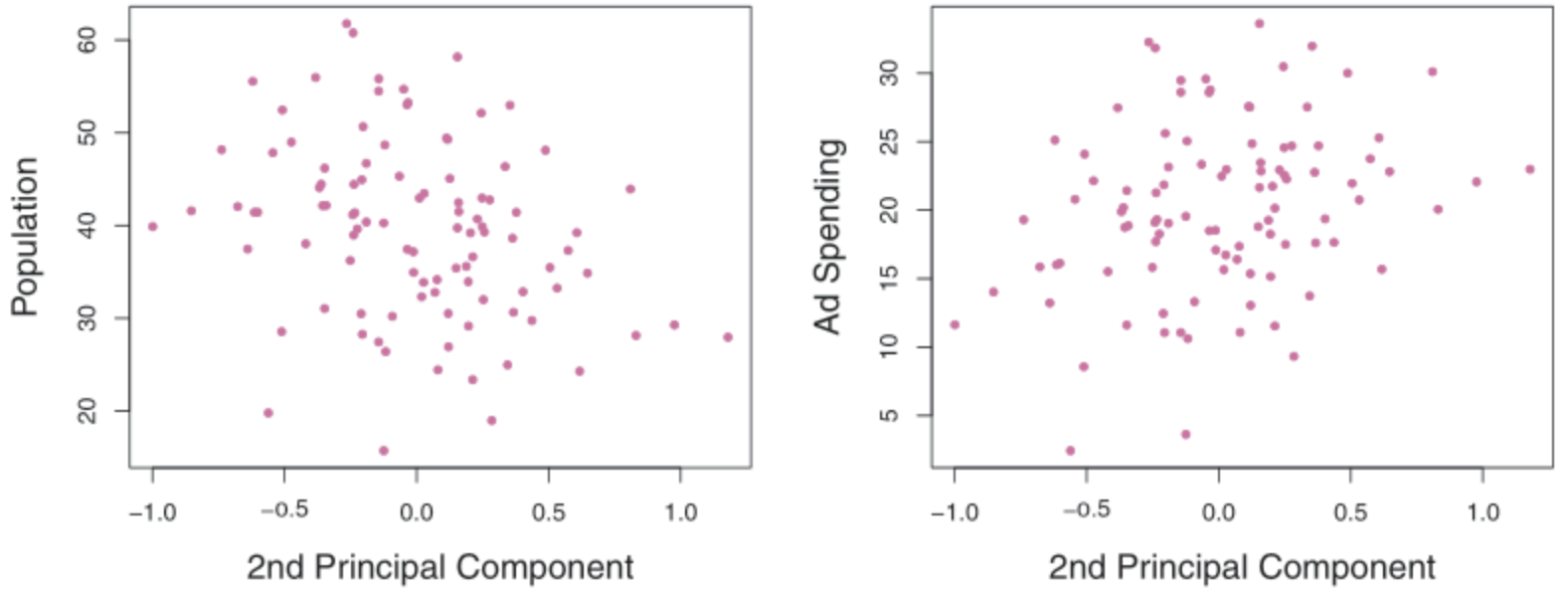

FIGURE 6.17. *Plots of the second principal component scores z_{i2} versus* pop *and* ad*. The relationships are weak.*

Source: ISL

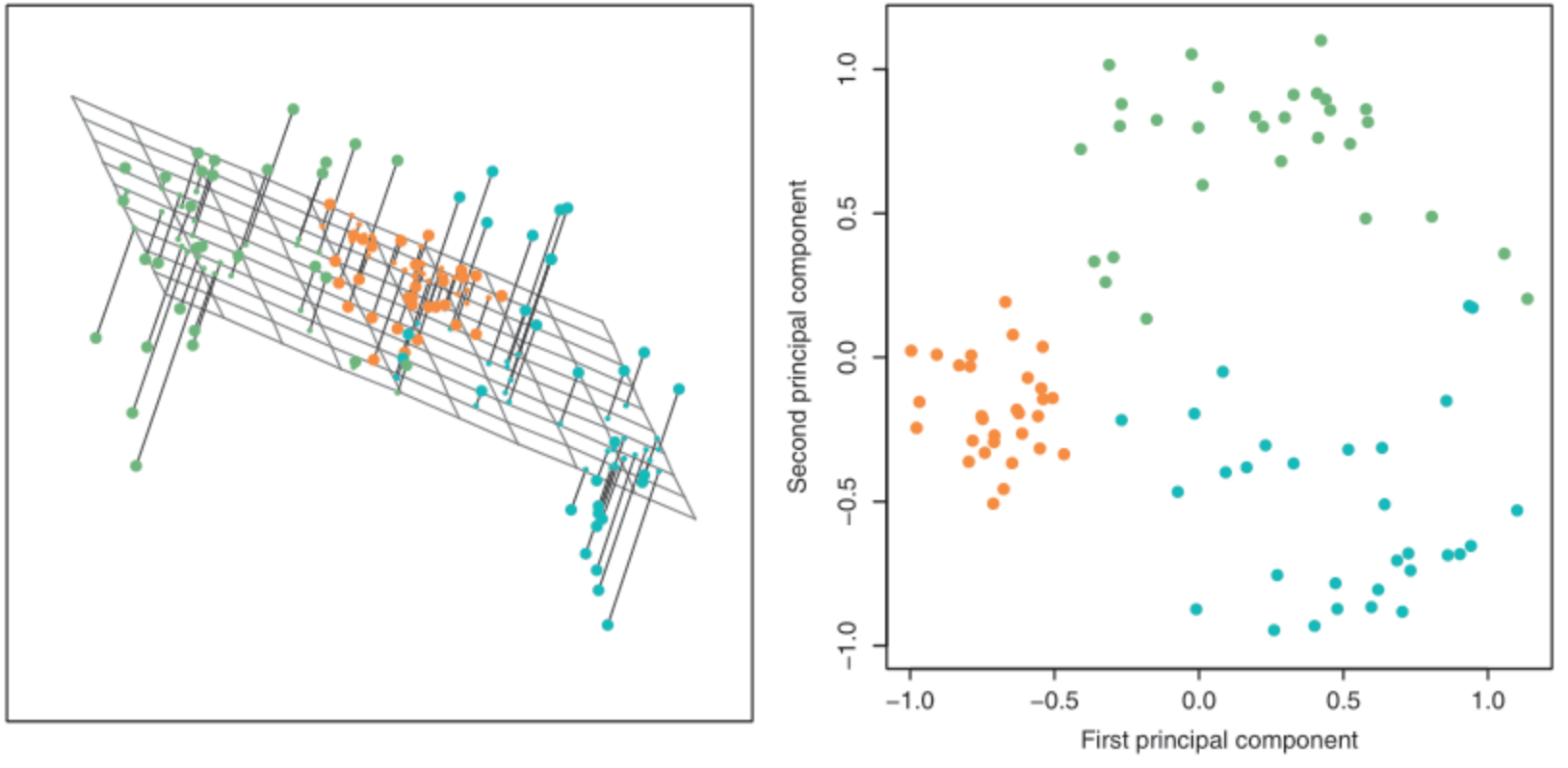

FIGURE 10.2. *Ninety observations simulated in three dimensions.* Left: *the first two principal component directions span the plane that best fits the data. It minimizes the sum of squared distances from each point to the plane.* Right: *the first two principal component score vectors give the coordinates of the projection of the 90 observations onto the plane. The variance in the plane is maximized.*

Source: ISL

Some PCA issues

- **Scaling the variables**: The results depend on the scale of the variables. If the variables are on very different scales, apply PCA after rescaling them by dividing with their SDs.
- **Non-uniqueness of PC**: The PC are not unique if there are ties among the eigenvalues. Even if there are no such ties, each PC (and hence the corresponding score vector) is unique only up to a sign flip. This is because a PC specifies a direction in the p-dimensional space, and flipping the signs of all its elements does not change the direction.
- **Proportion of variance explained**: The PC Z_k has variance λ_k and the total variance is $\sum_{k=1}^{p} \lambda_k$. Therefore, the PVE by Z_k is

$$\text{PVE}(k) = \frac{\lambda_k}{\sum_{l=1}^{p} \lambda_l}.$$

  This is a non-increasing function of k.

- **How many PCs to use?** We would like to use the smallest number of PCs that give us a *good* understanding of the data but this number depends on the application. Make a **scree plot** — the plot of PVE(k) against k — and look for the *elbow*, the point beyond which PVE levels off. May also plot the cumulative PVE — $\sum_{k=1}^{M} \text{PVE}(k)$ — against M. Usually, $M \ll p$.
- **Maximum # PC**: p — if $\text{rank}(\mathbf{X}) = r < p$, implying that $\mathbf{S}$ is singular, we have only r PC.
- **What if $p \geq n$?** Everything works except the maximum # PC is $\min\{n-1, p\}$ rather than p.
- **Singular value decomposition (SVD)**: Can also use an SVD of $\mathbf{X}$ instead of the spectral decomposition of $\mathbf{S}$ to form the PC — generally more numerically stable and also used by the software

Back to Dimension Reduction Methods

Set up: n obs on response Y and predictors $X_1, \ldots, X_p$

Principal components regression (PCR): Compute first M PCs $Z_1, \ldots, Z_M$ of the original p predictors and use the PCs as predictors in a linear model that is fit by least squares ($M \ll p$)

Implicit assumption: The directions in which the predictors vary the most are also the directions that are associated with Y

- Choose M by cross-validation
- Equivalent to least squares if $M = p$
- Standardize the predictors before performing PCR
- Does not perform variable selection
- Closely related to ridge regression
- The PCs are computed in an **unsupervised** way, i.e., Y is not used to help find the PC directions. Therefore, it may not work well if the above assumption does not hold.

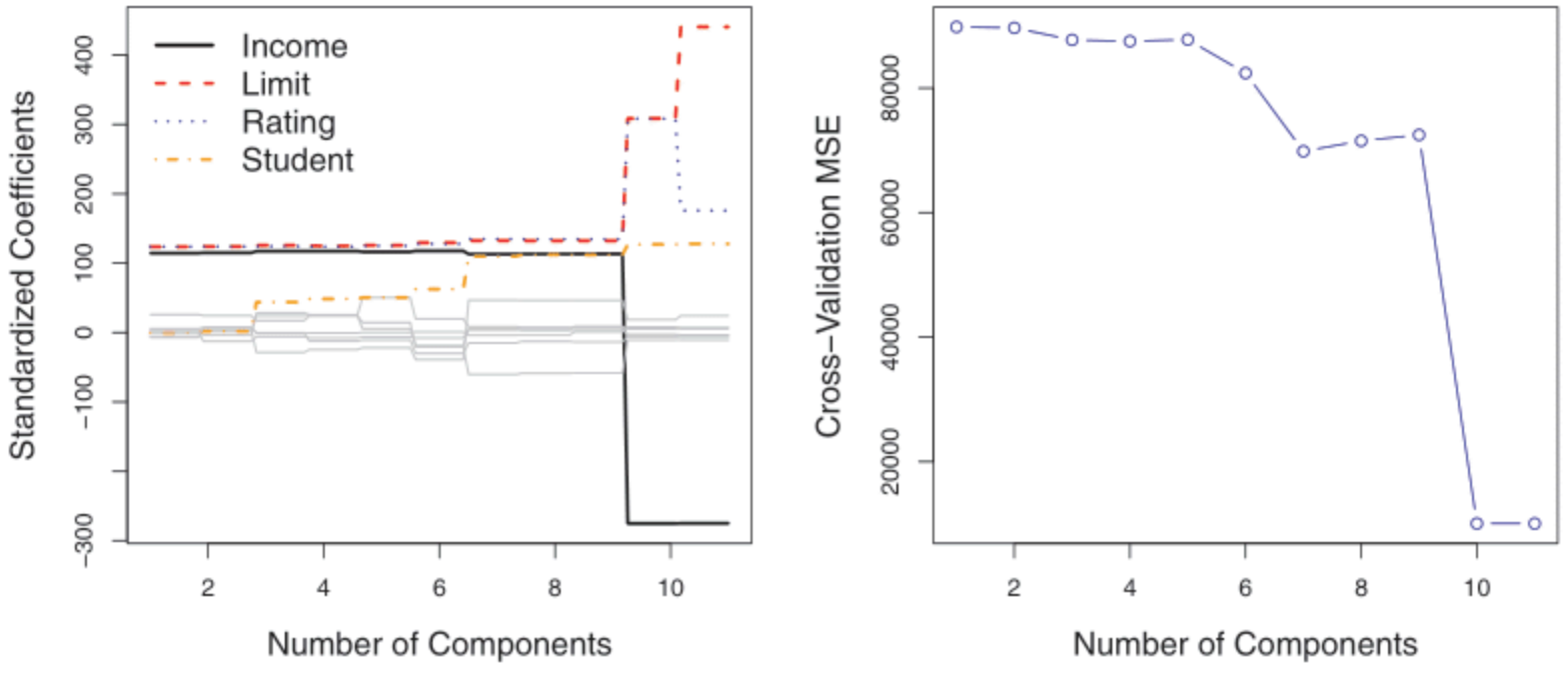

FIGURE 6.20. Left: *PCR standardized coefficient estimates on the* `Credit` *data set for different values of* M. Right: *The ten-fold cross validation MSE obtained using PCR, as a function of* M.

Source: ISL

Partial Least Squares (PLS)

PLS is a **supervised** alternative to PCR in which Y is used to determine the PC directions. The resulting $Z_1, \ldots, Z_M$ not only approximate the original predictors well but are also related to Y. Thus, in a sense, PLS attempts to find directions that help explain *both* response and the predictors. It works in two steps after standardizing the predictors.

Step 1: Find the first M PLS directions $Z_1, \ldots, Z_M$ in a supervised way.

- Compute $Z_1 = \phi_{11} X_1 + \ldots + \phi_{p1} X_p$, where ϕ_{j1} is the slope of simple linear regression of Y and X_j. Since this slope is proportional to the correlation between Y and X_j, the highest weight is placed on the predictors that are most strongly related to Y.

- To compute Z_2, first regress each X_j on Z_1 and compute the residuals r_j. These residuals provide the remaining information that has not been explained the first PLS direction. Compute Z_2 using the residuals r_j in the same way as Z_1 is computed using the X_j.
- Repeat to identify the PLS directions $Z_3, \ldots, Z_M$.

Step 2: Use least squares to fit a linear model to Y using $Z_1, \ldots, Z_M$ as predictors.

- Choose M by cross-validation
- Equivalent to least squares when $M = p$
- Does not perform variable selection
- Behaves much like PCR and ridge regression
- Supervised dimension reduction conducted by PLS can reduce bias but may also potentially increase variance, so its overall benefit over PCR may not be clear

Optimization Problems Solved by PCA and PLS

Recall that the kth PC direction ϕ_k solves:

$$\max_a \ \text{var}(a^T X),$$

subject to

$$a^T a = 1, \ a^T S \phi_k = 0, \ l = 1, \ldots, k-1.$$

Here $\text{var}(a^T X) = a^T \mathbf{S} a$. Whereas, the kth PLS direction ϕ_k solves

$$\max_a \ \text{corr}^2(\mathbf{Y}, \mathbf{X}a) \, \text{var}(a^T X)$$

subject to the same constraints as PCA. Typically the variance aspect dominates, making PLS similar to PCR

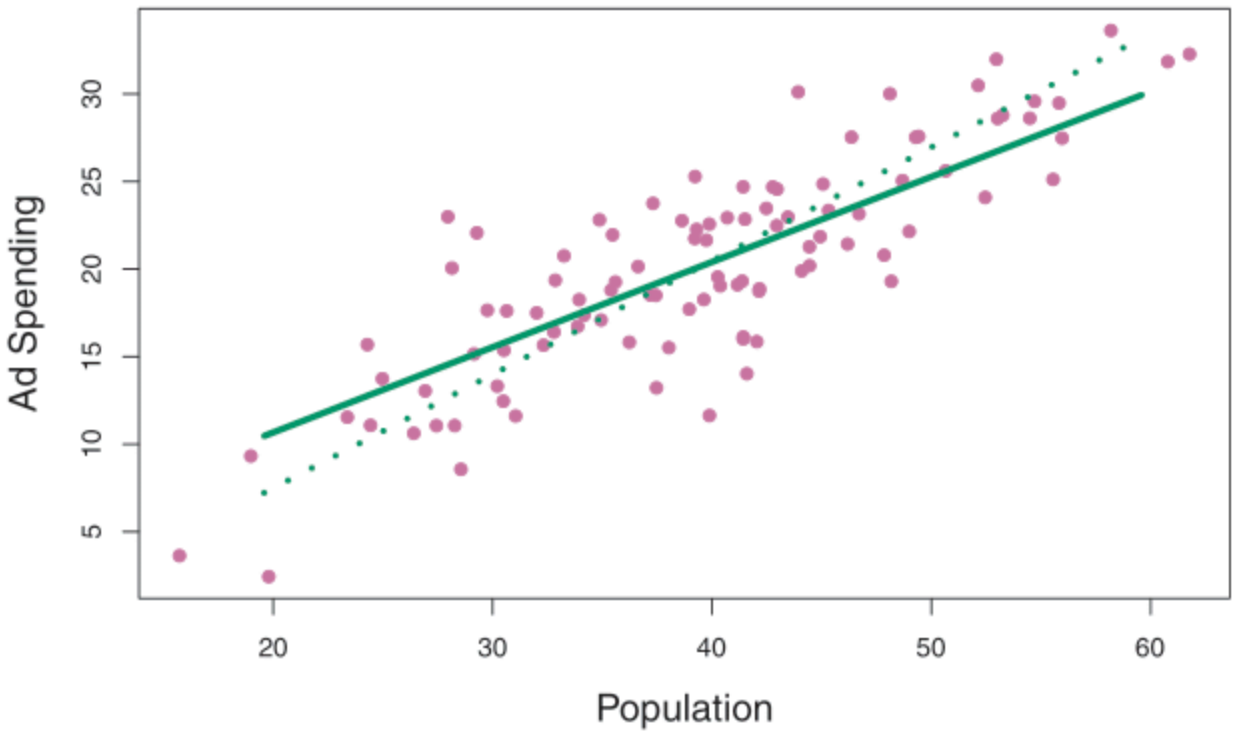

FIGURE 6.21. *For the advertising data, the first PLS direction (solid line) and first PCR direction (dotted line) are shown.*

Source: ISL

Some Considerations in High Dimensions

High dimensional data: $p \geq n$, often $p \gg n$

- Least squares method does not work as it gives a perfect fit
- Use regularization methods, with appropriately selected tuning parameter
- Should not use traditional measures of model fit on training data, including R^2, C_p, AIC, BIC, p-values, etc.
- Instead, report results on an independent test set or test errors estimated by cross-validation
- **Multicollinearity**: When $p > n$, any predictor can be written as a linear combination of all of the other predictors, i.e., multicollinearity exists for sure. Therefore, we need to be careful in interpreting results.
- In the best case scenario, what we get is *one of many possible models*, which must be further validated on independent datasets.

Dimension Reduction Methods - Additionals

Python applications

- Principal Component Regression
- Partial Least Squares