CPSC 8420 Advanced Machine Learning
Principal Component Analysis

Complexity of Least Square Solution

We have derived that for Multiple Linear Regression model:

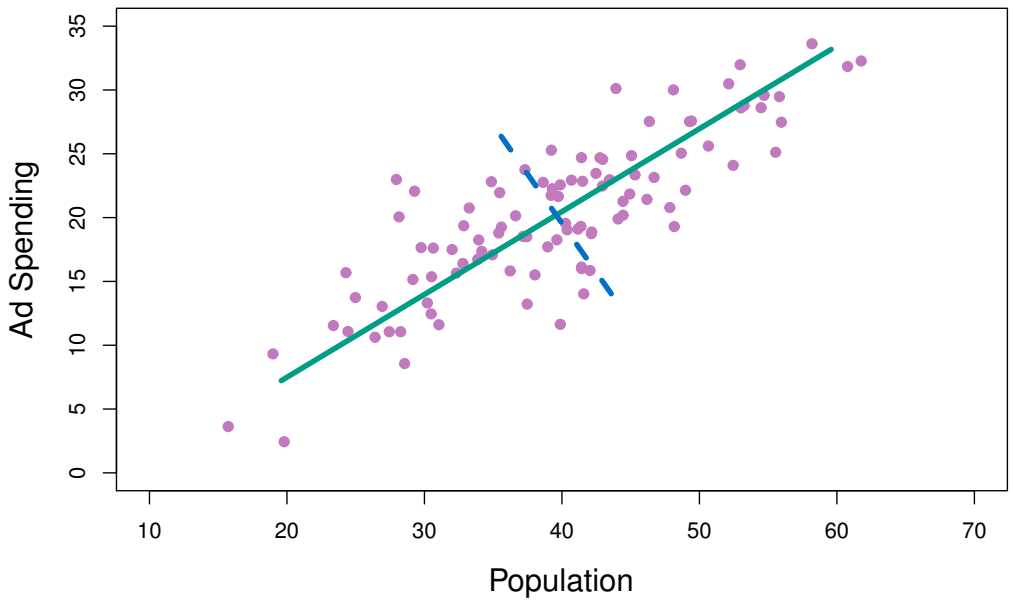
$$\min \|Ax - y\|^2 \tag{1}$$

where $A \in \mathbb{R}^{n \times (p+1)}$, the solution is $(A^T A)^{-1} A^T y$. Noting that $A^T A \in \mathbb{R}^{(p+1) \times (p+1)}$, which will bring $\mathcal{O}(p^3)$ complexity for the inversion operation. When p is large, it becomes computationally demanding. For example, assume we can reduce $p = 100$ to 10, then we can save the time consumption to $\frac{1}{1000}$ as before. Therefore, dimension reduction is valuable in real-world if only it won't lose important information of original data.

Principal Components Analysis

- PCA produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
- PCA is a tool for data reduction.
- Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

How to formulate Principal Component Analysis



- 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 Analysis

- The first principal component of a set of features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the features

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p \tag{2}$$

that has the largest variance. By normalized, we mean that

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1 \tag{3}$$

Principal Components Analysis

- The first principal component of a set of features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the features

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p \tag{2}$$

that has the largest variance. By normalized, we mean that

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1 \tag{3}$$

- We refer to the elements $\phi_{11}, \ldots, \phi_{p1}$ as the loadings of the first principal component; together, the loadings make up the principal component loading vector, $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})^T$.

Principal Components Analysis

- The first principal component of a set of features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the features

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p \tag{2}$$

that has the largest variance. By normalized, we mean that

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1 \tag{3}$$

- We refer to the elements $\phi_{11}, \ldots, \phi_{p1}$ as the loadings of the first principal component; together, the loadings make up the principal component loading vector, $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})^T$.

- We constrain the loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

Computation

- Suppose we have a $n \times p$ data set $\mathbf{X}$. Since we are only interested in variance, we assume that each of the variables in $\mathbf{X}$ has been centered to have mean zero.

Computation

- Suppose we have a $n \times p$ data set $\mathbf{X}$. Since we are only interested in variance, we assume that each of the variables in $\mathbf{X}$ has been centered to have mean zero.
- We then look for the linear combination of the sample feature values of the form

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \ldots \phi_{p1} x_{ip} \tag{4}$$

for $i = 1, \ldots, n$ that has largest sample variance, subject to the constraint that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. $z_{11}, z_{21}, \ldots, z_{n1}$ are referred as scores of the first principal component.

Computation

- Suppose we have a $n \times p$ data set $\mathbf{X}$. Since we are only interested in variance, we assume that each of the variables in $\mathbf{X}$ has been centered to have mean zero.
- We then look for the linear combination of the sample feature values of the form

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \ldots \phi_{p1} x_{ip} \tag{4}$$

  for $i = 1, \ldots, n$ that has largest sample variance, subject to the constraint that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. $z_{11}, z_{21}, \ldots, z_{n1}$ are referred as scores of the first principal component.
- The underlying optimization problem can be solved via a singular-value decomposition of the matrix $\mathbf{X}$, a standard technique in linear algebra.

Deriving the solution

$$\max_{\phi_{11},\phi_{21},\ldots,\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{j1}x_{ij})^2 \quad s.t. \quad \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

denote that

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \in \mathbb{R}^{n\times p}$$

then we can reformulate as:

$$\max_{\phi_1} \|X\phi_1\|^2 \quad s.t. \quad \sum_{j=1}^{p}\phi_1^2 = 1$$

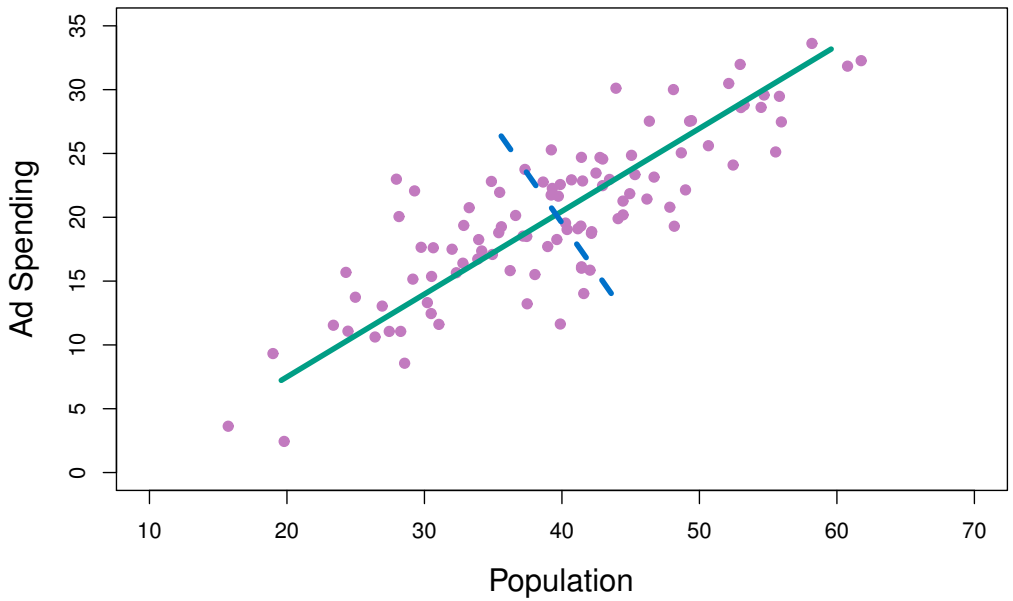
which is equivalent to:

$$\max_{\phi_1} trace(\phi_1^T X^T X \phi_1) \quad s.t. \quad \sum_{j=1}^{p}\phi_1^2 = 1$$

Deriving the solution

If we compute $[U, S] = svd(X^T X)$, then we have $\phi_1 = U(:, 1)$, and the variance $\|X\phi_1\|^2 = S(1, 1)$, similarly we have $\phi_k = U(:, k)$ and $\|X\phi_k\|^2 = S(k, k)$.

Ad Spending & Population



$$Z_1 = 0.839 \times (pop - \bar{pop}) + 0.544 \times (ad - \bar{ad})$$

$$z_{i1} = 0.839 \times (pop_i - \bar{pop}) + 0.544 \times (ad_i - \bar{ad}) \qquad (5)$$

Here $\phi_{11} = 0.839$ and $\phi_{21} = 0.544$ are the *principal component loadings*, which define the direction referred to green line.
$Z_{11}, Z_{21}, \ldots, Z_{n1}$ are known as *principal component scores*.

Ad Spending & Population

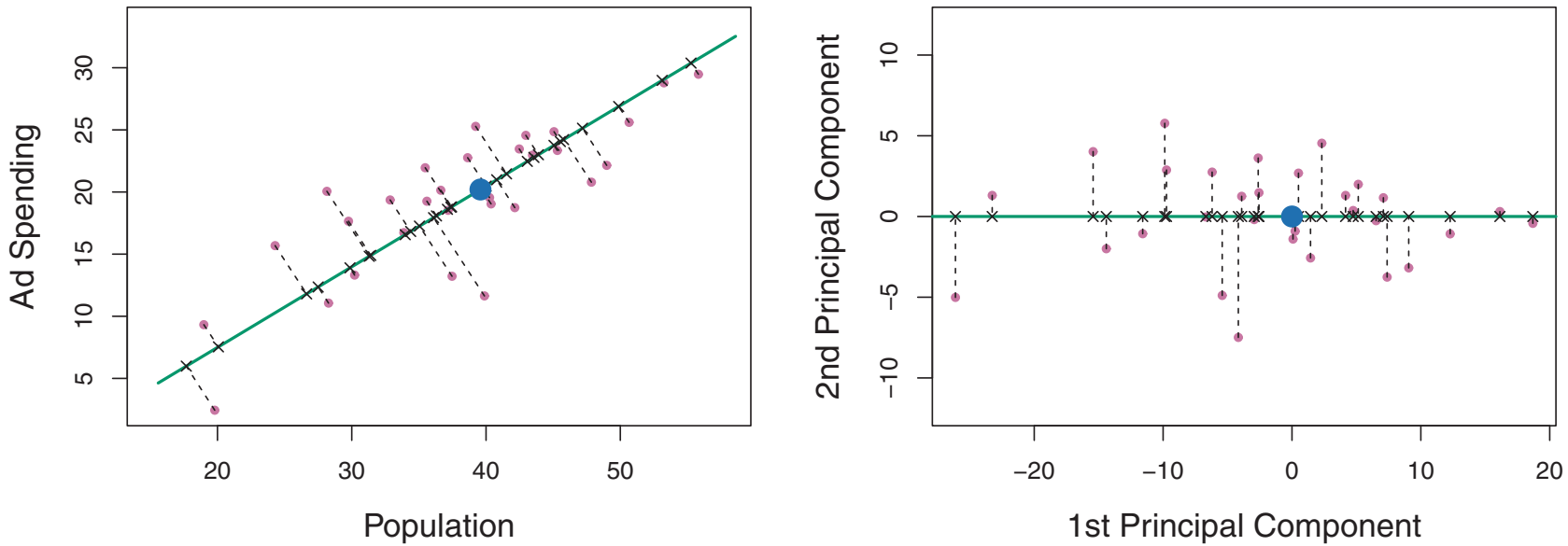


FIGURE 6.15. *A subset of the advertising data. The mean* pop *and* ad *budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all* n *of the observations.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

Ad Spending & Population

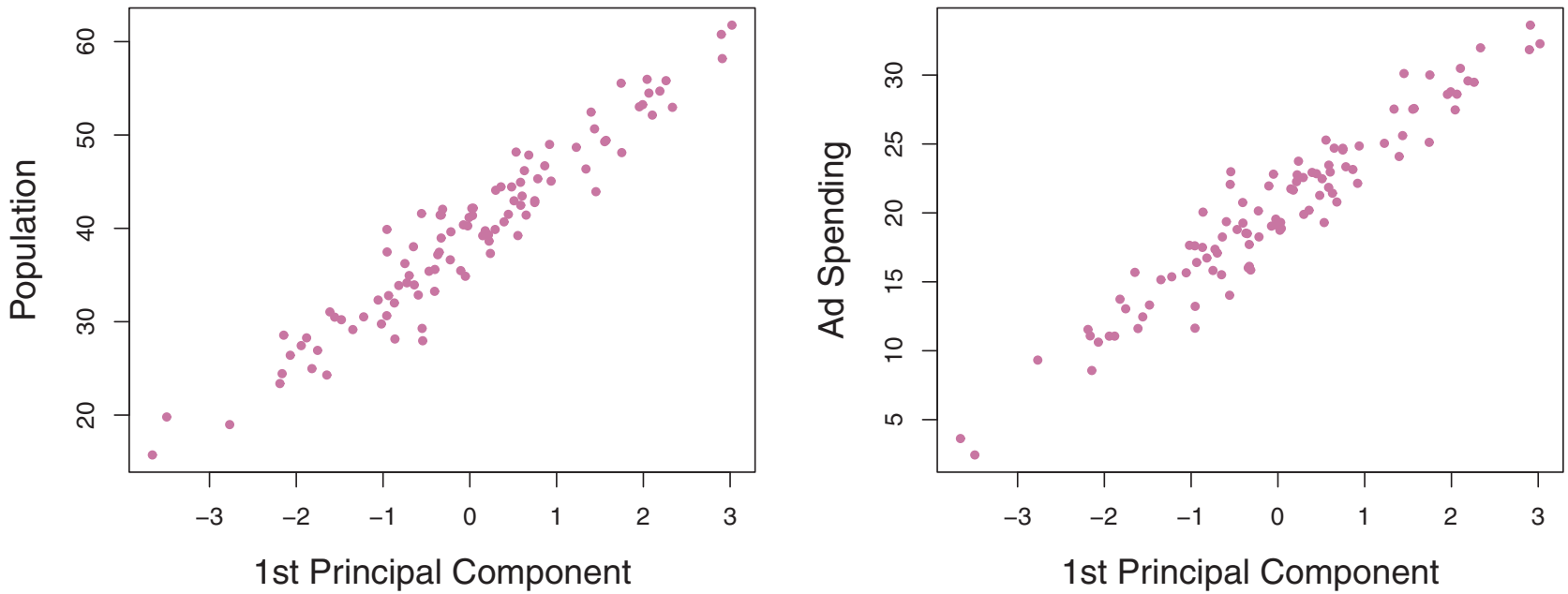
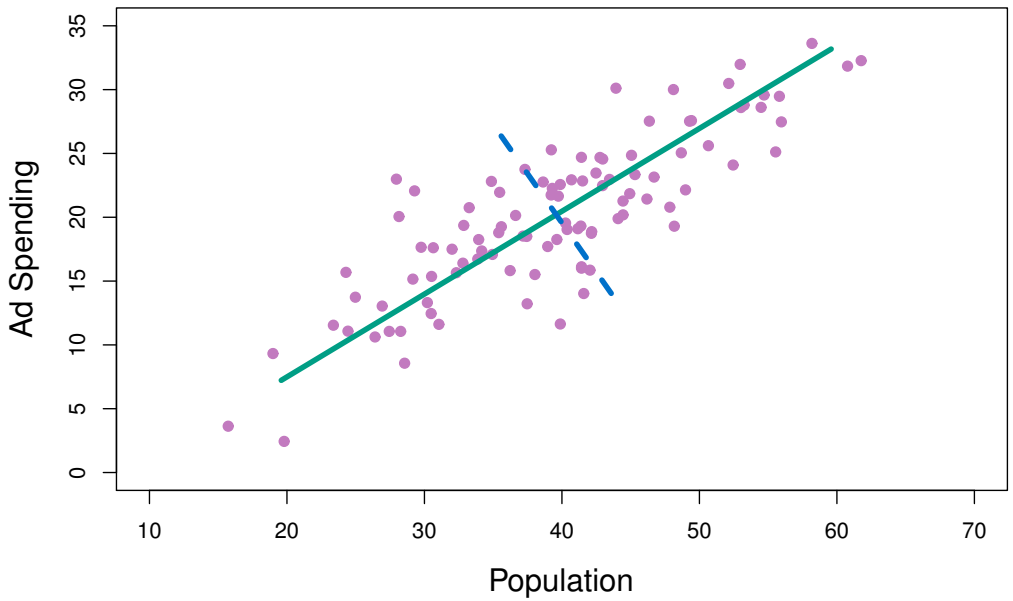


FIGURE 6.16. *Plots of the first principal component scores z_{i1} versus* pop *and* ad. *The relationships are strong.*

Ad Spending & Population



$$Z_2 = 0.544 \times (pop - \bar{pop}) - 0.839 \times (ad - \bar{ad})$$

$$z_{i2} = 0.544 \times (pop_i - \bar{pop}) - 0.839 \times (ad_i - \bar{ad}) \quad (6)$$

Ad Spending & Population

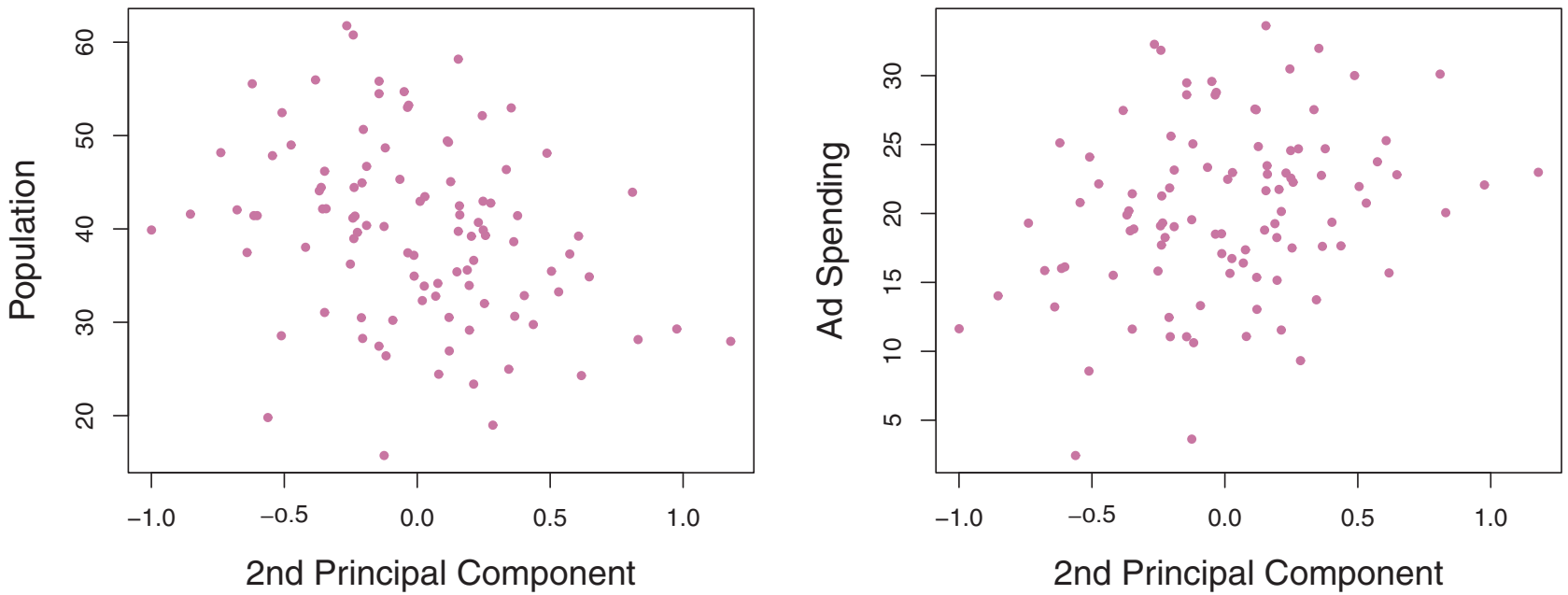


FIGURE 6.17. *Plots of the second principal component scores z_{i2} versus* pop *and* ad*. The relationships are weak.*

Example

- *USAarrests data*: For each of the fifty states in the US, the data set contains the number of arrests per 100,000 residents for each of three crimes: *Assault*, *Murder*, and *Rape*. We also record *UrbanPop*, the percent of the population in each state living in urban areas.
- The principal component score vectors have length $n = 50$, and the principal component loading vectors have length $p = 4$.
- PCA was performed after standardizing each variable to have mean zero and standard deviation one.

Example

The principal component loading vectors ϕ_1 and ϕ_2.

	PC1	**PC2**
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

Example

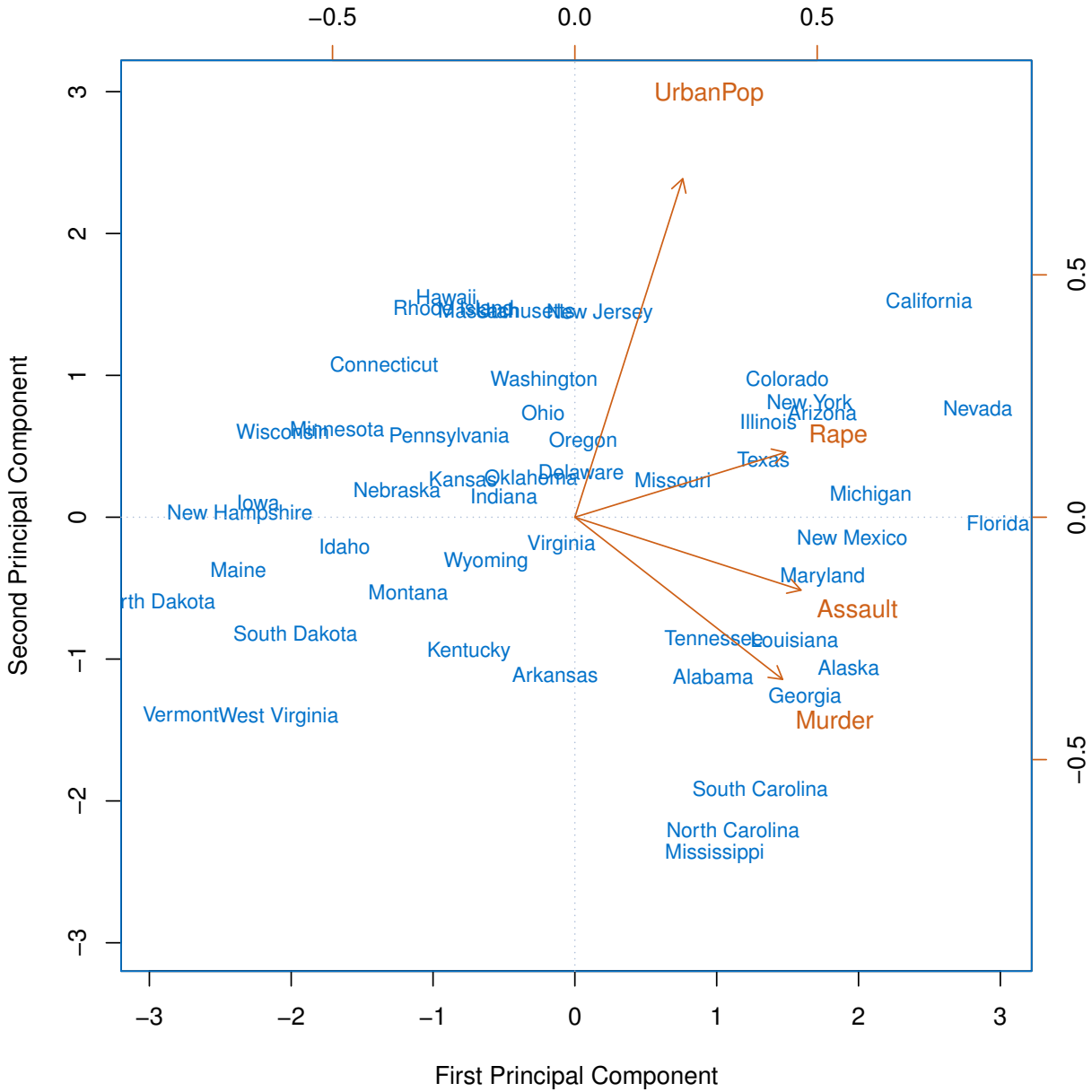

UrbanPop
Rape
Assault
Murder
California
New Jersey
Connecticut
Washington
Colorado
New York
Arizona
Nevada
Illinois
Ohio
Wisconsin
Minnesota
Pennsylvania
Oregon
Texas
Delaware
Missouri
Nebraska
Kansas
Oklahoma
Indiana
Michigan
Iowa
New Hampshire
Florida
Idaho
Virginia
New Mexico
Wyoming
Maine
Maryland
Montana
South Dakota
Tennessee
Louisiana
Kentucky
Alaska
Arkansas
Alabama
Georgia
Vermont
West Virginia
South Carolina
North Carolina
Mississippi
Second Principal Component
First Principal Component
−0.5
0.0
0.5
−3
−2
−1
0
1
2
3

Figure details

- The figure is known as a biplot, because it displays both the principal component scores and the principal component loadings.
- The blue state names represent the scores for the first two principal components.
- The orange arrows indicate the first two principal component loading vectors (with axes on the top and right). For example, the loading for Murder on the first component is 0.54, and its loading on the second principal component -0.42 [the word "Murder" is centered at the point (0.54, -0.42)].

Scaling is Important

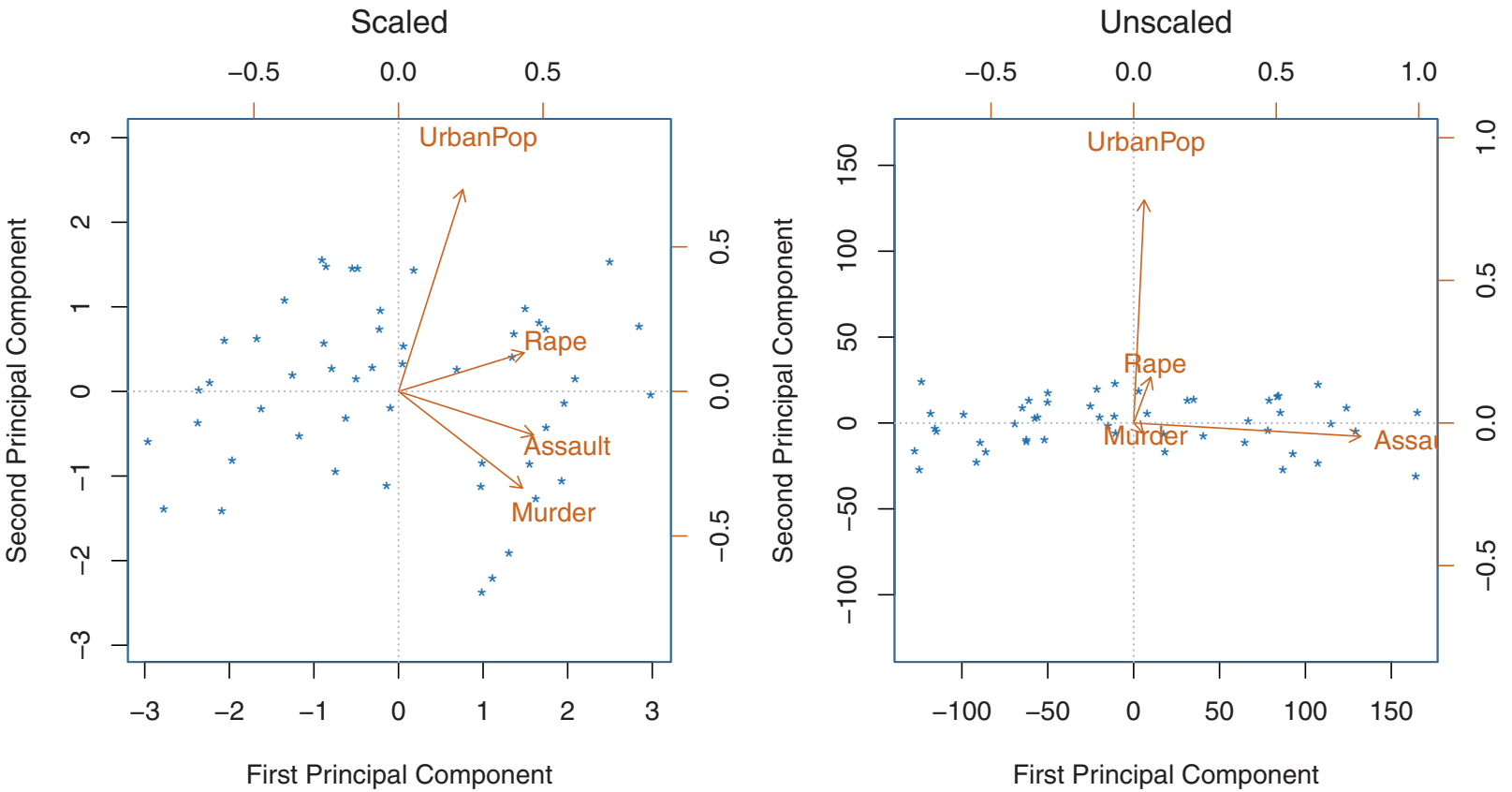
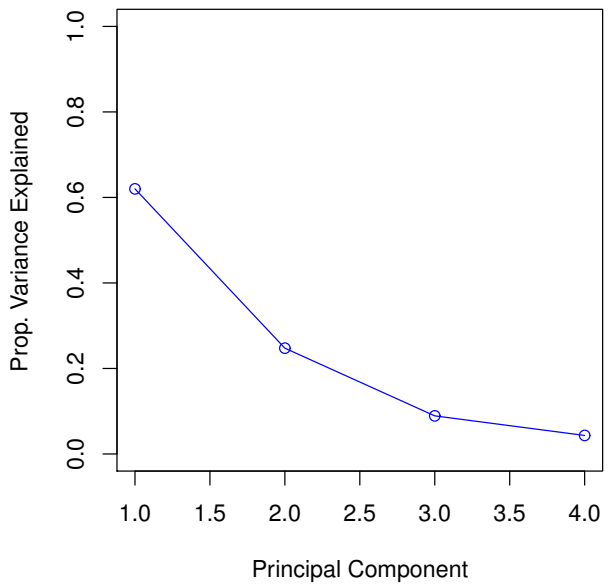


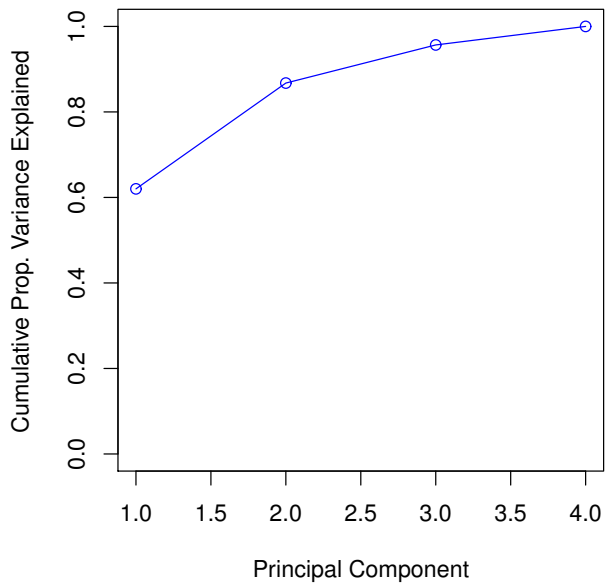
FIGURE 10.3. *Two principal component biplots for the* `USArrests` *data.* Left: *the same as Figure 10.1, with the variables scaled to have unit standard deviations.* Right: *principal components using unscaled data.* `Assault` *has by far the largest loading on the first principal component because it has the highest variance among the four variables. In general, scaling the variables to have standard deviation one is recommended.*

How many principal components should we use?

If we use principal components as a summary of our data, how many components are sufficient?

- No simple answer to this question, as cross-validation is not available for this purpose.
- Data scientists often use so called "scree plots" as a guide and look for an "elbow"





Another perspective to interpret PCA

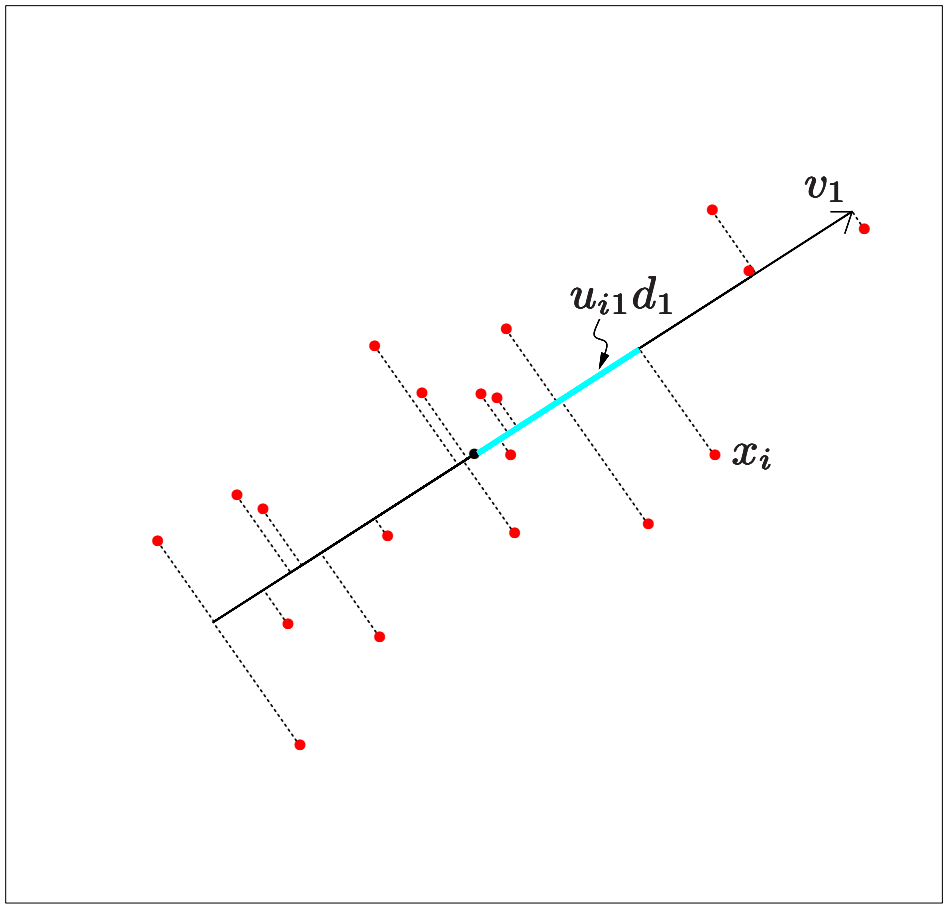


FIGURE 14.20. *The first linear principal component of a set of data. The line minimizes the total squared distance from each point to its orthogonal projection onto the line.*

Formulation

Denote the observations by $x_1, x_2, \ldots, x_N$, and consider the rank-q linear model for representing them:

$$f(\lambda) = \mu + \boldsymbol{V}_q \lambda$$

where μ is a location vector in $\mathbb{R}^p$, $\boldsymbol{V}_q$ is a $p \times q$ matrix with q orthogonal unit vectors as columns, and λ is a q vector of parameters. Fitting such a model to the data by least squares amounts to minimizing the *reconstruction error*:

$$\min_{\mu, \lambda_i, \boldsymbol{V}_q} \sum_{i=1}^{N} \|x_i - \mu - \boldsymbol{V}_q \lambda_i\|^2$$

Optimization

We can partially optimize for μ and the λ_i to obtain:

$$\hat{\mu} = \bar{x}$$
$$\hat{\lambda} = \boldsymbol{V}_q^T (x_i - \bar{x})$$

which is equivalent to find orthogonal matrix $\boldsymbol{V}_q$:

$$\begin{aligned} \min_{\boldsymbol{V}_q} \sum_{i=1}^{N} \|(x_i - \bar{x}) - \boldsymbol{V}_q \boldsymbol{V}_q^T (x_i - \bar{x})\|^2 \\ = \|X - \boldsymbol{V}_q \boldsymbol{V}_q^T X\|_F^2 \end{aligned}$$

where $X \in \mathbb{R}^{p \times N}$

Another perspective to interpret PCA

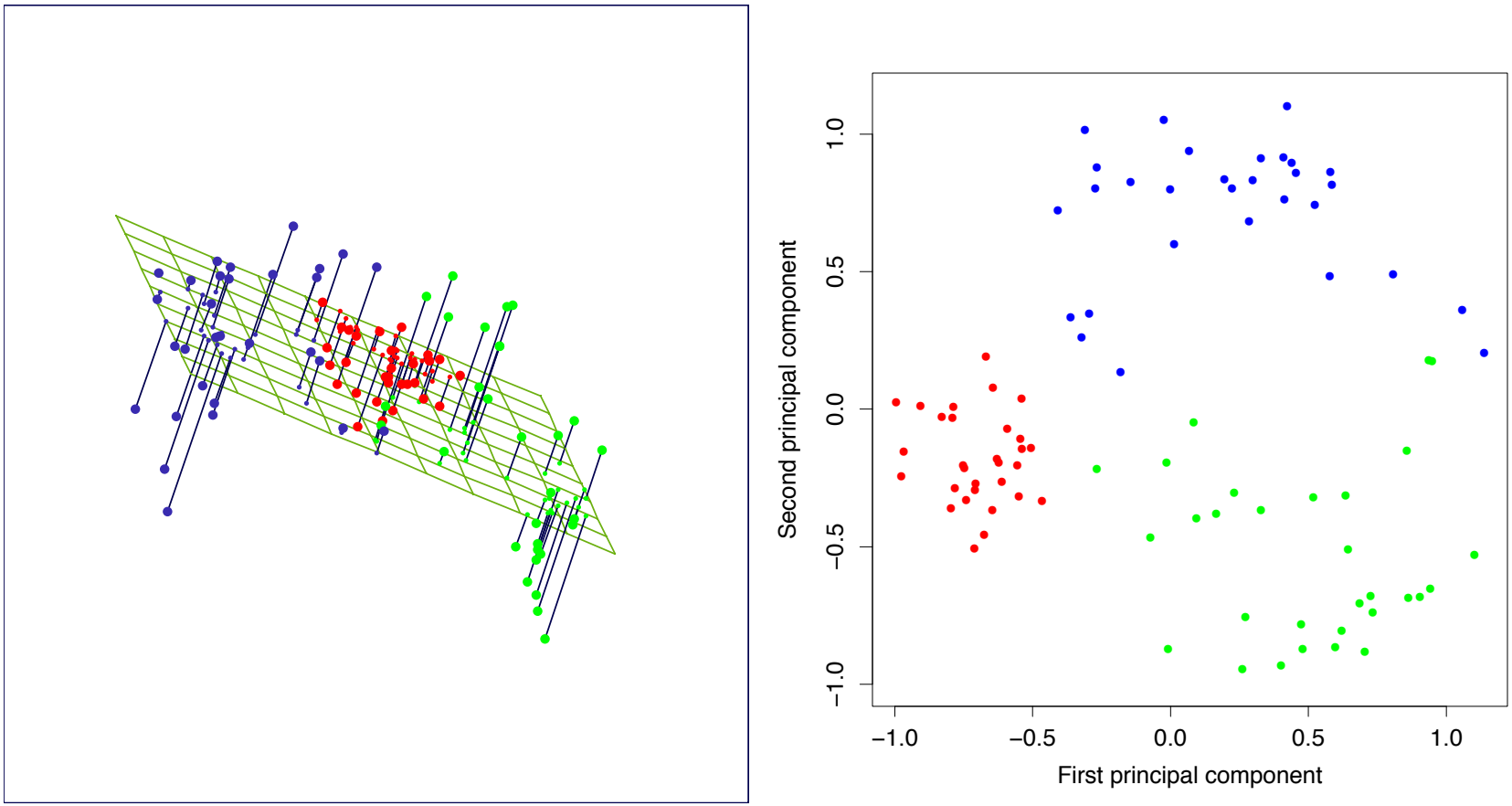


FIGURE 14.21. *The best rank-two linear approximation to the half-sphere data. The right panel shows the projected points with coordinates given by* $\mathbf{U}_2\mathbf{D}_2$*, the first two principal components of the data.*

Hand-written threes

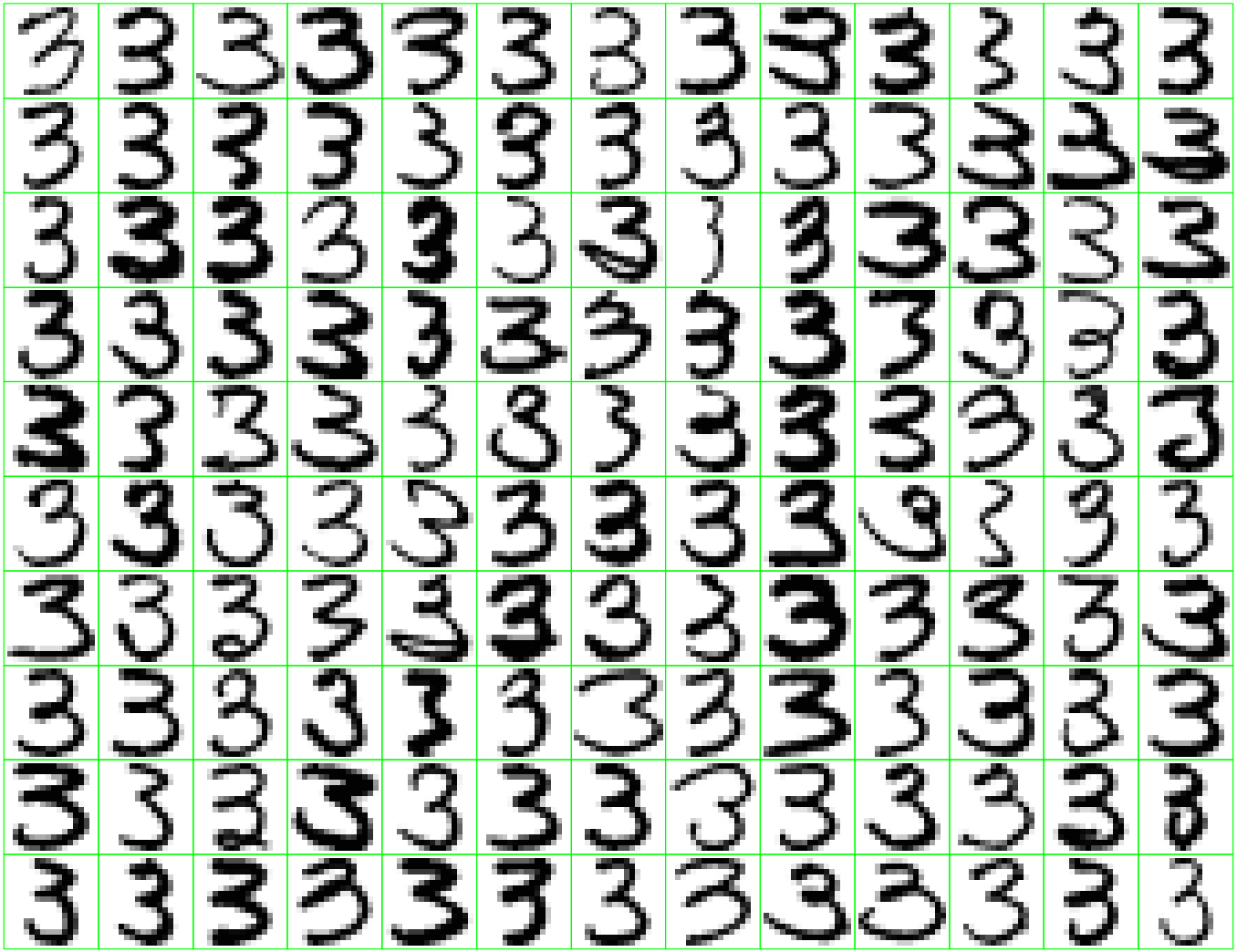
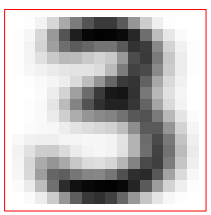


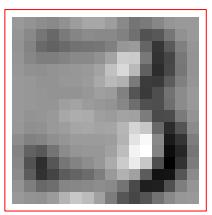
FIGURE 14.22. *A sample of* 130 *handwritten 3's shows a variety of writing styles.*

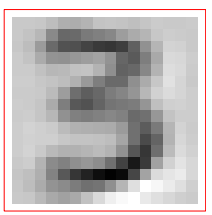
Hand-written threes

$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$$

$$= \bar{x} + \lambda_1 \cdot v_1 + \lambda_2 \cdot v_2.$$







Hand-written threes