Principal Component Analysis

Lu Wang

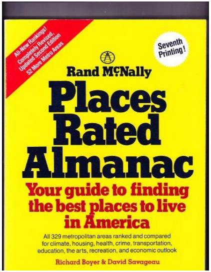
Department of Statistical Science

03/29/2016

Example: Places Rated

Boyer and Savageau rated 329 communities according to 9 criteria:

1. Climate
2. Housing Cost
3. Health Care
4. Crime Rate
5. Transportation
6. Education
7. The Arts
8. Recreation
9. Economics



Question: Are all those criteria super important? Or are some more important than others?

Learning Objectives

Principal component analysis (PCA) is a data reduction technique.

- reduce the number of possibly correlated variables of interest into a smaller set of uncorrelated components
- a useful tool in exploratory data analysis and for predictive modeling

Outline:

1. Understand PCA procedure
2. Assess how many principal components should be considered in an analysis

Learning Objectives

Principal component analysis (PCA) is a data reduction technique.

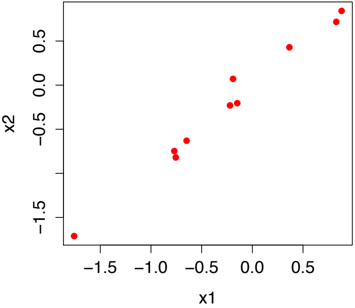
- reduce the number of possibly correlated variables of interest into a smaller set of uncorrelated components
- a useful tool in exploratory data analysis and for predictive modeling

Outline:

1. Understand PCA procedure
2. Assess how many principal components should be considered in an analysis

Why It May Be Possible to Reduce Dimensions?

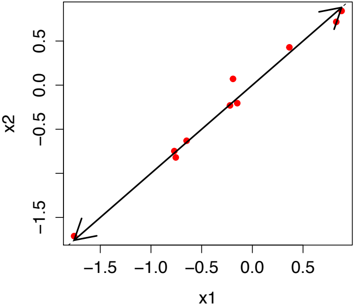
When we have correlations between the variables, the data may more or less fall on a line or plane in a lower number of dimensions.



This line could be used as a new axis (one-dimensional) to represent the variation among data points.

Why It May Be Possible to Reduce Dimensions?

When we have correlations between the variables, the data may more or less fall on a line or plane in a lower number of dimensions.



This line could be used as a new axis (one-dimensional) to represent the variation among data points.

PCA Procedure

Question: Can we reduce the large number of correlated variables to a few uncorrelated linear combinations of them?

Suppose that we have a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ with $p \times p$ variance-covariance matrix Σ.

Consider the linear combinations

$$
\begin{aligned}
Y_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\
Y_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\
&\vdots \\
Y_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p
\end{aligned}
$$

- Let $\mathbf{e}_i = (e_{i1}, \ldots, e_{ip})'$, $i = 1, 2, \ldots, p$.

$$\operatorname{var}(Y_i) = \mathbf{e}_i'\Sigma\mathbf{e}_i \qquad \operatorname{cov}(Y_i, Y_j) = \mathbf{e}_i'\Sigma\mathbf{e}_j$$

PCA Procedure

Question: Can we reduce the large number of correlated variables to a few uncorrelated linear combinations of them?

Suppose that we have a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ with $p \times p$ variance-covariance matrix Σ.

Consider the linear combinations

$$\begin{aligned} Y_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ Y_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ &\vdots \\ Y_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned}$$

- Let $\mathbf{e}_i = (e_{i1}, \ldots, e_{ip})'$, $i = 1, 2, \ldots, p$.

$$\mathrm{var}(Y_i) = \mathbf{e}_i'\Sigma\mathbf{e}_i \qquad \mathrm{cov}(Y_i, Y_j) = \mathbf{e}_i'\Sigma\mathbf{e}_j$$

PCA Procedure

Question: Can we reduce the large number of correlated variables to a few uncorrelated linear combinations of them?

Suppose that we have a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ with $p \times p$ variance-covariance matrix Σ.

Consider the linear combinations

$$\begin{aligned} Y_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ Y_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ &\vdots \\ Y_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned}$$

- Let $\mathbf{e}_i = (e_{i1}, \ldots, e_{ip})'$, $i = 1, 2, \ldots, p$.

$$\text{var}(Y_i) = \mathbf{e}_i'\Sigma\mathbf{e}_i \qquad \text{cov}(Y_i, Y_j) = \mathbf{e}_i'\Sigma\mathbf{e}_j$$

PCA Procedure

Question: Can we reduce the large number of correlated variables to a few uncorrelated linear combinations of them?

Suppose that we have a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ with $p \times p$ variance-covariance matrix Σ.

Consider the linear combinations

$$\begin{aligned}
Y_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\
Y_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\
&\vdots \\
Y_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p
\end{aligned}$$

- Let $\mathbf{e}_i = (e_{i1}, \ldots, e_{ip})'$, $i = 1, 2, \ldots, p$.

$$\mathrm{var}(Y_i) = \mathbf{e}_i'\Sigma\mathbf{e}_i \qquad \mathrm{cov}(Y_i, Y_j) = \mathbf{e}_i'\Sigma\mathbf{e}_j$$

Principal Components

- ***First Principal Component (PC1):*** Y_1
  - PC1 is the linear combination of X-variables that has maximum variance among all linear combinations.
  - Select $\mathbf{e}_1 = (e_{11}, e_{12}, \ldots, e_{1p})'$ that maximizes

$$\text{var}(Y_1) = \mathbf{e}_1' \Sigma \mathbf{e}_1 \quad s.t. \quad \mathbf{e}_1' \mathbf{e}_1 = 1.$$

- ***Second Principal Component (PC2):*** Y_2
  - PC2 is the linear combination of X-variables that accounts for as much of the remaining variation as possible
  - the correlation between PC1 and PC2 is 0.
  - select $\mathbf{e}_2 = (e_{21}, e_{22}, \ldots, e_{2p})'$ that maximizes $\text{var}(Y_2) = \mathbf{e}_2' \Sigma \mathbf{e}_2$

$$\text{s.t. } \mathbf{e}_2' \mathbf{e}_2 = 1 \quad \text{and} \quad \text{cov}(Y_1, Y_2) = \mathbf{e}_1' \Sigma \mathbf{e}_2 = 0.$$

Principal Components

- ***First Principal Component (PC1):*** Y_1
  - PC1 is the linear combination of X-variables that has maximum variance among all linear combinations.
  - Select $\mathbf{e}_1 = (e_{11}, e_{12}, \ldots, e_{1p})'$ that maximizes

$$\text{var}(Y_1) = \mathbf{e}_1' \Sigma \mathbf{e}_1 \quad s.t. \quad \mathbf{e}_1' \mathbf{e}_1 = 1.$$

- ***Second Principal Component (PC2):*** Y_2
  - PC2 is the linear combination of X-variables that accounts for as much of the remaining variation as possible
  - the correlation between PC1 and PC2 is 0.
  - select $\mathbf{e}_2 = (e_{21}, e_{22}, \ldots, e_{2p})'$ that maximizes $\text{var}(Y_2) = \mathbf{e}_2' \Sigma \mathbf{e}_2$

$$\text{s.t. } \mathbf{e}_2' \mathbf{e}_2 = 1 \quad \text{and} \quad \text{cov}(Y_1, Y_2) = \mathbf{e}_1' \Sigma \mathbf{e}_2 = 0.$$

Principal Components

All subsequent principal components have the same properties:

- linear combinations that account for as much of the remaining variation as possible
- not correlated with the other principal components

Question: How do we obtain the coefficients e_{ij} ?

Eigenvalue decomposition for Real Symmetric Matrices

Every $p \times p$ real symmetric matrix Σ can be decomposed as

$$\Sigma = Q \Lambda Q^T$$

where Λ is a $p \times p$ diagonal matrix whose entries are the eigenvalues of Σ and Q is a $p \times p$ orthogonal matrix whose columns are the corresponding eigenvectors.

Principal Components

All subsequent principal components have the same properties:

- linear combinations that account for as much of the remaining variation as possible
- not correlated with the other principal components

Question: How do we obtain the coefficients e_{ij} ?

Eigenvalue decomposition for Real Symmetric Matrices

Every $p \times p$ real symmetric matrix Σ can be decomposed as

$$\Sigma = Q \Lambda Q^T$$

where Λ is a $p \times p$ diagonal matrix whose entries are the eigenvalues of Σ and Q is a $p \times p$ orthogonal matrix whose columns are the corresponding eigenvectors.

Solution of coefficients e_{ij}

- Let λ_1 through λ_p denote the eigenvalues of the variance-covariance matrix Σ which are ordered such that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$$

- Let vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ denote the corresponding orthogonal eigenvectors. Then $\mathbf{e}_i$ will be the coefficients of i^{th} principal component, $i = 1, 2, \ldots, p$.

Discuss:

1. Why are all the principal components obtained in this way uncorrelated with one another?
2. What is the variance for i^{th} principal component, $i = 1, \ldots, p$?
3. * Why this is a valid solution?

Solution of coefficients e_{ij}

- Let λ_1 through λ_p denote the eigenvalues of the variance-covariance matrix Σ which are ordered such that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$$

- Let vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ denote the corresponding orthogonal eigenvectors. Then $\mathbf{e}_i$ will be the coefficients of i^{th} principal component, $i = 1, 2, \ldots, p$.

Discuss:

1. Why are all the principal components obtained in this way uncorrelated with one another?
2. What is the variance for i^{th} principal component, $i = 1, \ldots, p$?
3. * Why this is a valid solution?

Solution of coefficients e_{ij}

- Let λ_1 through λ_p denote the eigenvalues of the variance-covariance matrix Σ which are ordered such that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$$

- Let vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ denote the corresponding orthogonal eigenvectors. Then $\mathbf{e}_i$ will be the coefficients of i^{th} principal component, $i = 1, 2, \ldots, p$.

Discuss:

1. Why are all the principal components obtained in this way uncorrelated with one another?
2. What is the variance for i^{th} principal component, $i = 1, \ldots, p$?
3. * Why this is a valid solution?

Interpretation of Principal Components (I)

- The variance for the i^{th} principal component is

$$\mathrm{var}(Y_i) = \mathrm{var}(\mathbf{e}_i' X_i) = \mathbf{e}_i' \Sigma \mathbf{e}_i = \lambda_i, \;\; i = 1, 2, \ldots, p.$$

- The total variation of $\mathbf{X}$ is defined as the trace of Σ.

$$\begin{aligned} \mathrm{trace}(\Sigma) &= \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2 \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p \end{aligned}$$

- The i^{th} principal component explains the following proportion of the total variation:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Interpretation of Principal Components (I)

- The variance for the i^{th} principal component is

$$\mathrm{var}(Y_i) = \mathrm{var}(\mathbf{e}_i' X_i) = \mathbf{e}_i' \Sigma \mathbf{e}_i = \lambda_i, \ \ i = 1, 2, \ldots, p.$$

- The total variation of $\mathbf{X}$ is defined as the trace of Σ.

$$\begin{aligned} \mathrm{trace}(\Sigma) &= \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2 \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p \end{aligned}$$

- The i^{th} principal component explains the following proportion of the total variation:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Interpretation of Principal Components (I)

- The variance for the i^{th} principal component is

$$\text{var}(Y_i) = \text{var}(\mathbf{e}_i' X_i) = \mathbf{e}_i' \Sigma \mathbf{e}_i = \lambda_i, \;\; i = 1, 2, \ldots, p.$$

- The total variation of $\mathbf{X}$ is defined as the trace of Σ.

$$\begin{aligned} \text{trace}(\Sigma) &= \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2 \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p \end{aligned}$$

- The i^{th} principal component explains the following proportion of the total variation:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

How many PCs should be considered?

- A related quantity is the proportion of variation explained by the first k principal components:

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

- If this quantity is large, not much information is lost by considering only the first k principal components.

Eigenvalues of $\hat{\Sigma}$ in the Places Rated data (Boyer and Savageau)

PC	Eigen	Prop	Cumu
1	0.3775	0.7227	0.7227
2	0.0511	0.0977	0.8204
3	0.0279	0.0525	0.8739
4	0.0230	0.0440	0.9178
5	0.0168	0.0321	0.9500
6	0.0120	0.0229	0.9728
7	0.0085	0.0162	0.9890
8	0.0039	0.0075	0.9966
9	0.0018	0.0034	1.0000
Total	0.5225		

How many PCs should be considered?

- A related quantity is the proportion of variation explained by the first k principal components:

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

- If this quantity is large, not much information is lost by considering only the first k principal components.

Eigenvalues of $\hat{\Sigma}$ in the Places Rated data (Boyer and Savageau)

PC	Eigen	Prop	Cumu
1	0.3775	0.7227	0.7227
2	0.0511	0.0977	0.8204
3	0.0279	0.0525	0.8739
4	0.0230	0.0440	0.9178
5	0.0168	0.0321	0.9500
6	0.0120	0.0229	0.9728
7	0.0085	0.0162	0.9890
8	0.0039	0.0075	0.9966
9	0.0018	0.0034	1.0000
Total	0.5225		

How many PCs should be considered?

- A related quantity is the proportion of variation explained by the first k principal components:

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

- If this quantity is large, not much information is lost by considering only the first k principal components.

Eigenvalues of $\hat{\Sigma}$ in the Places Rated data (Boyer and Savageau)

PC	Eigen	Prop	Cumu
1	0.3775	0.7227	0.7227
2	0.0511	0.0977	0.8204
3	0.0279	0.0525	0.8739
4	0.0230	0.0440	0.9178
5	0.0168	0.0321	0.9500
6	0.0120	0.0229	0.9728
7	0.0085	0.0162	0.9890
8	0.0039	0.0075	0.9966
9	0.0018	0.0034	1.0000
Total	0.5225		

How many PCs should be considered?

- An alternative method is to look at a Scree Plot – The number of components can be determined at the point beyond which the remaining eigenvalues are all relatively small.