PENNSTATE

1855

STAT 505 - Applied Multivariate Statistical

Start Here!

- Welcome to STAT 505!
- Faculty login (PSU Access Account)

Lessons

- Lesson 0: Matrices and Vectors
- Lesson 1: Graphical Display of Multivariate Data
- Lesson 2: Measures of Central Tendency, Dispersion and Association
- Lesson 3: Linear Combinations of Random Variables
- Lesson 4: Multivariate Normal Distribution
- Lesson 5: Sample Mean Vector and Sample Correlation and Related Inference Problems
- Lesson 6: Multivariate Conditional Distribution and Partial Correlation
- Lesson 7: Principal Components Analysis (PCA)
  - 7.1 - Principal Component Analysis (PCA) Procedure
  - 7.2 - How do we find the coefficients?
  - 7.3 - Example: Places Rated
  - 7.4 - Interpretation of the Principal Components
  - 7.5 - Alternative: Standardize the Variables
  - 7.6 - Example: Places Rated after Standardization
  - 7.7 - Once the Components Have Been Calculated
  - 7.8 - Summary
- Lesson 8: Canonical Correlation Analysis
- Lesson 9: Factor Analysis
- Lesson 10: Discriminant Analysis
- Lesson 11: Inferences Regarding Multivariate Population Mean
- Lesson 12: Cluster Analysis
- Lesson 13: Multivariate Analysis of Variance (MANOVA)

Home // Lesson 7: Principal Components Analysis (PCA)

7.2 - How do we find the coefficients?

Printer-friendly version

How do we find the coefficients e_{ij} for a principal component?

The solution involves the eigenvalues and eigenvectors of the variance-covariance matrix Σ.

Solution:

We are going to let λ_1 through λ_p denote the eigenvalues of the variance-covariance matrix Σ. These are ordered so that λ_1 has the largest eigenvalue and λ_p is the smallest.

$$\lambda_1 \le \lambda_2 \le \dots \le \lambda_p$$

We are also going to let the vectors $\mathbf{e}_1$ through $\mathbf{e}_p$

$$\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_p$$

denote the corresponding eigenvectors. It turns out that the elements for these eigenvectors will be the coefficients of our principal components.

The variance for the *i*th principal component is equal to the *i*th eigenvalue.

$$\mathbf{var}(Y_i) = \text{var}(e_{i1}X_1 + e_{i2}X_2 + \dots e_{ip}X_p) = \lambda_i$$

Moreover, the principal components are uncorrelated with one another.

$$\text{cov}(Y_i, Y_j) = 0$$

The variance-covariance matrix may be written as a function of the

- Lesson 14: Repeated Measures Analysis

Resources

- SAS Tour
- SAS Interactive Data Analysis
- A Quick Introduction to Minitab
- Worked Examples from the Course That Use Software

eigenvalues and their corresponding eigenvectors. This is determined by using the Spectral Decomposition Theorem. This will become useful later when we investigate topics under factor analysis.

Spectral Decomposition Theorem

The variance-covarianc matrix can be written as the sum over the p eigenvalues, multiplied by the product of the corresponding eigenvector times its transpose as shown in the first expression below:

$$\begin{aligned}\Sigma &= \sum_{i=1}^{p}\lambda_i \mathbf{e}_i\mathbf{e}'_i\\ &\cong \sum_{i=1}^{k}\lambda_i \mathbf{e}_i\mathbf{e}'_i\end{aligned}$$

The second expression is a useful approximation if $\lambda_{k+1}, \lambda_{k+2}, \dots, \lambda_p$ are small. We might approximate Σ by

$$\sum_{i=1}^{k}\lambda_i \mathbf{e}_i\,\mathbf{e}'_i$$

Again, this will become more useful when we talk about factor analysis.

Earlier in the course we defined the total variation of **X** as the trace of the variance-covariance matrix, or if you like, the sum of the variances of the individual variables. This is also equal to the sum of the eigenvalues as shown below:

$$\begin{aligned}trace(\Sigma) &= \sigma^2_1 + \sigma^2_2 + \dots + \sigma^2_p\\ &= \lambda_1 + \lambda_2 + \dots + \lambda_p\end{aligned}$$

This will give us an interpretation of the components in terms of the amount of the full variation explained by each component. The proportion of variation explained by the ith principal component is then going to be defined to be the eigenvalue for that component divided by the sum of the eigenvalues. In other words, the ith principal component explains the following proportion of the total variation:

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

A related quantity is the proportion of variation explained by the first k principal component. This would be the sum of the first k eigenvalues divided by its total variation.

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

Naturally, if the proportion of variation explained by the first k principal components is large, then not much information is lost by considering only the first k principal components.

Why It May Be Possible to Reduce Dimensions

When we have correlations (multicollinarity) between the x-variables, the data may more or less fall on a line or plane in a lower number of dimensions. For instance, imagine a plot of two x-variables that have a nearly perfect correlation. The data points will fall close to a straight line. That line could be used as a new (one-dimensional) axis to represent the variation among data points. As another example, suppose that we have verbal, math, and total SAT scores for a sample of students. We have three variables, but really (at most) two dimensions to the data because total= verbal +math, meaning the third variable is completely determined by the first two. The reason for saying “at most” two dimensions is that if there is a strong correlation between verbal and math, then it may be possible that there is only one true dimension to the data.

Note

All of this is defined in terms of the population variance-covariance matrix Σ which is unknown. However, we may estimate Σ by the sample variance-covariance matrix which is given in the standard formula here:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{x}})(\mathbf{X}_i - \bar{\mathbf{x}})'$$

Procedure

Compute the eigenvalues $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_p$ of the sample variance-covariance matrix **S**, and the corresponding eigenvectors $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \dots, \hat{\mathbf{e}}_p$.

Then we will define our estimated principle components using the eigenvectors as our coefficients:

$$\begin{array}{lll}\hat{Y}_1 & = & \hat{e}_{11}X_1 + \hat{e}_{12}X_2 + \cdots + \hat{e}_{1p}X_p \\ \hat{Y}_2 & = & \hat{e}_{21}X_1 + \hat{e}_{22}X_2 + \cdots + \hat{e}_{2p}X_p \\ & & \vdots \\ \hat{Y}_p & = & \hat{e}_{p1}X_1 + \hat{e}_{p2}X_2 + \cdots + \hat{e}_{pp}X_p \end{array}$$

Generally, we only retain the first *k* principal component. Here we must balance two conflicting desires:

1. To obtain the simplest possible interpretation, we want *k* to be as small as possible. If we can explain most of the variation just by two principal components then this would give us a much simpler description of the data. The smaller *k* is the smaller amount of variation is explained by the first *k* component.

2. To avoid loss of information, we want the proportion of variation explained by the first *k* principal components to be large. Ideally as close to one as possible; i.e., we want

$$\frac{\hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_k}{\hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p} \cong 1$$

‹ 7.1 - Principal Component Analysis (PCA) Procedure up 7.3 - Example: Places Rated ›

Printer-friendly version

Drupal

© 2014 The Pennsylvania State University. All rights reserved.