High Dimensional Data Analysis

Cluster Analysis

Cluster Analysis

- Introduction
- Distance and similarity functions
- Hierarchical clustering
- Dynamic clustering

Introduction

- Cluster analysis is a multivariate statistical method for classifying samples or indicators (variables).
- Cluster analysis methods are categorized into Q-type clustering and R-type clustering based on the classification objective. Q-type clustering clusters samples, while R-type clustering clusters variables (indicators).
- This chapter focuses on Q-type clustering methods for clustering samples, including hierarchical clustering and dynamic clustering methods. The latter primarily utilizes the k-means clustering method, which is widely used in big data analysis.

Distance and similarity functions

- **Clustering of samples**: A common method is to regard each sample as a point in multidimensional space and establish distance in space. Points with closer distances are classified into the same category, and points with farther distances are classified into different categories.
- **Clustering of variables**: usually the similarity coefficient between variables is calculated, and variables with larger similarity coefficients are classified into the same category, while variables with smaller similarity coefficients are classified into different categories.

The above distance and similarity coefficient are generally for quantitative data. For qualitative data, the distance between samples or the similarity coefficient between variables needs to be defined according to the specific data structure.

Distance and similarity functions

Definition

For sample points $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^t$ and $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \cdots, x_{jp})^t$, distances and correlation coefficients are defined as:

Clustering Statistics

- **Euclidean Distance:**

$$d_{ij}(2) = \left[\sum_{k=1}^{p}(x_{ik} - x_{jk})^2\right]^{\frac{1}{2}}$$

- **Mahalanobis Distance:**

$$d_{ij}(M) = (\boldsymbol{x}_i - \boldsymbol{x}_j)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)$$

Distance and similarity functions

Clustering Statistics

- **Lance–Williams Distance:**

$$d_{ij}(LW) = \frac{1}{p}\sum_{k=1}^{p}\frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}$$

- **Correlation Coefficient:**

$$r_{ij} = \frac{\sum_{k=1}^{p}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{p}(x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{p}(x_{jk} - \bar{x}_j)^2}}$$

Hierarchical clustering

The basic idea of hierarchical clustering

First, each sample is considered as a single point class, that is, each sample is a class. Then, the classes are merged one by one, and the two classes with the smallest distance are merged together each time. After merging, the distance between the classes is recalculated. This process continues until all samples are classified into one class, and the process is drawn into a systematic cluster diagram.

Steps of Hierarchical Clustering Method

1. Calculate the distances between n samples to obtain the distance matrix $D^{(0)}$, which is an $n \times n$ symmetric matrix.
2. Select the smallest off-diagonal element D_{kl} in $D^{(0)}$, and merge clusters G_k and G_l into a new cluster denoted as G_m, i.e., $G_m = G_k \cup G_l$.
3. Calculate the distances between the new cluster G_m and other clusters to obtain a new distance matrix $D^{(1)}$.
4. Repeat the above two steps for $D^{(1)}$ as done for $D^{(0)}$, obtaining $D^{(2)}$, and continue this process until all elements are merged into one large cluster.
5. Draw the dendrogram (hierarchical clustering diagram).
6. Determine the number of clusters and the members of each cluster.

Single Linkage Method (Nearest neighbor)

Definition

The distance between two clusters is defined as the distance between their two closest elements. For clusters G_k and G_l, the distance is defined as:

$$D_{kl} = \min_{i \in G_k, j \in G_l} d_{ij}.$$

This clustering method is called the **Single Linkage Method**.

Recursive Formula

When clusters G_p and G_q merge to form a new cluster G_r, the distance between G_r and any other cluster G_k can be computed using the recursive formula:

$$D_{rk} = \min_{i \in G_r, j \in G_k} d_{ij} = \min\{D_{pk}, D_{qk}\}$$

Example 1: Single Linkage Clustering

Problem Statement

Suppose we have 5 one-dimensional samples with values 1, 2, 6.5, 7, and 11. Use the single linkage method to classify them.

Solution

Initially, each sample forms its own cluster. We use Euclidean distance and follow these steps:

1. Calculate the distances between the 5 clusters to obtain the initial inter-cluster distance matrix $D^{(0)}$.

	G_1	G_2	G_3	G_4	G_5
G_1	0.0				
G_2	1.0	0.0			
G_3	5.5	4.5	0.0		
G_4	6.0	5.0	0.5	0.0	
G_5	10.0	9.0	4.5	4.0	0.0

Example 1: Single Linkage Clustering

Step 2: Merge Closest Clusters

The smallest off-diagonal element in $D^{(0)}$ is $D_{34} = 0.5$. Therefore, merge G_3 and G_4 into a new cluster $G_6 = G_3 \cup G_4 = \{6.5, 7\}$. Calculate distances between G_6 and other clusters using the recursive formula. Obtain the new distance matrix $D^{(1)}$

Distance Matrix $D^{(1)}$

	G_1	G_2	G_5	G_6
G_1	0.0			
G_2	1.0	0.0		
G_5	10.0	9.0	0.0	
G_6	5.5	4.5	4.0	0.0

Where: $G_1 = \{1\}, G_2 = \{2\}, G_5 = \{11\}, G_6 = \{6.5, 7\}$

Example 1: Single Linkage Clustering

Step 3: Merge Next Closest Clusters

The smallest off-diagonal element in $D^{(1)}$ is $D_{12} = 1.0$. Therefore, merge G_1 and G_2 into a new cluster $G_7 = G_1 \cup G_2 = \{1, 2\}$. Calculate distances between G_7 and other clusters using the recursive formula. Obtain the new distance matrix $D^{(2)}$.

Distance Matrix $D^{(2)}$

	G_5	G_6	G_7
G_5	0.0		
G_6	4.0	0.0	
G_7	9.0	4.5	0.0

Where: $G_5 = \{11\}, G_6 = \{6.5, 7\}, G_7 = \{1, 2\}$

Distance Calculation

Example 1: Single Linkage Clustering

Step 4: Merge Next Closest Clusters

The smallest off-diagonal element in $D^{(2)}$ is $D_{56} = 4.0$. Therefore, merge G_5 and G_6 into a new cluster $G_8 = G_6 \cup G_5 = \{6.5, 7, 11\}$. Calculate distances between G_8 and other clusters using the recursive formula. Obtain the new distance matrix $D^{(3)}$

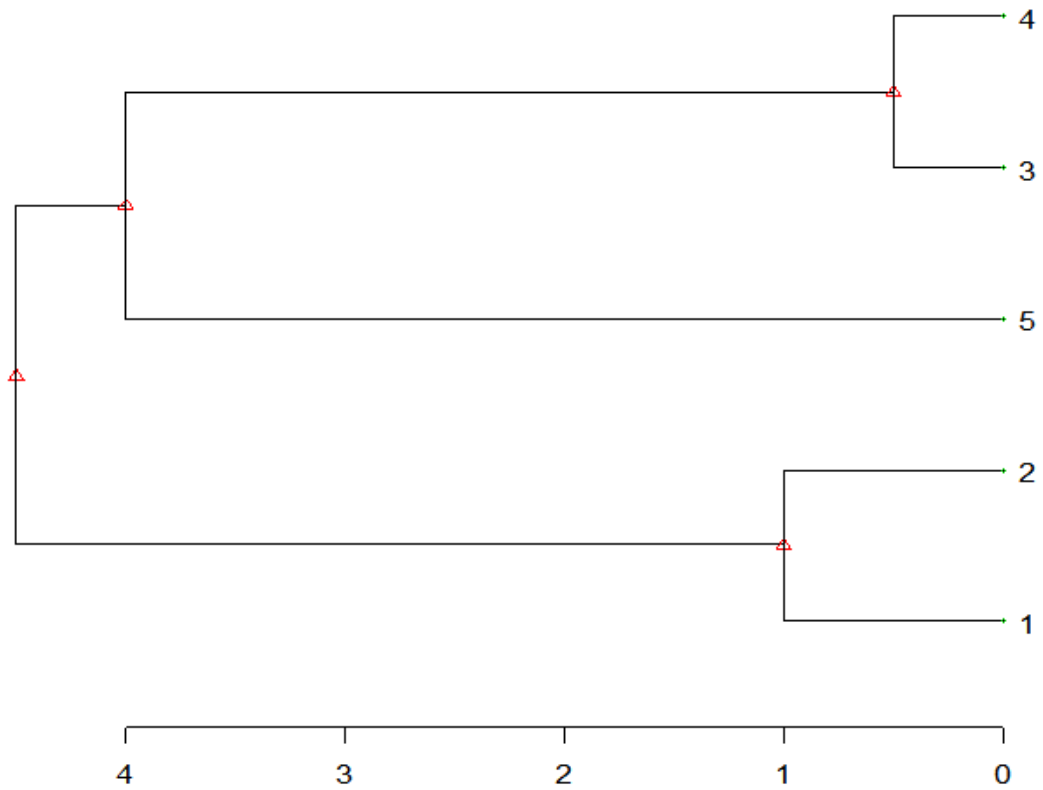
Distance Matrix $D^{(3)}$

	G_7	G_8
G_7	0.0	
G_8	4.5	0.0

Where: $G_7 = \{1, 2\}, G_8 = \{6.5, 7, 11\}$

Example 1: Single Linkage Clustering

- **If divided into 2 clusters:**
  - Cluster 1: $G_1 \cup G_2 = \{1, 2\}$
  - Cluster 2: $G_3 \cup G_4 \cup G_5 = \{6.5, 7, 11\}$
- **If divided into 3 clusters:**
  - Cluster 1: $G_1 \cup G_2 = \{1, 2\}$
  - Cluster 2: $G_5 = \{11\}$
  - Cluster 3: $G_3 \cup G_4 = \{6.5, 7\}$



Complete Linkage Method (Farthest Neighbor)

Definition

The distance between two clusters is defined as the distance between their two farthest elements. For clusters G_k and G_l, the distance is defined as:

$$D_{kl} = \max_{i \in G_k, j \in G_l} d_{ij}.$$

This clustering method is called the **Complete Linkage Method**.

Recursive Formula

When clusters G_p and G_q merge to form a new cluster G_r, the distance between G_r and any other cluster G_k can be computed using the recursive formula:

$$D_{rk} = \max_{i \in G_r, j \in G_k} d_{ij} = \max\{D_{pk}, D_{qk}\}$$

Example 2: Complete Linkage Clustering

Problem Statement

Suppose we have 5 one-dimensional samples with values 1, 2, 6.5, 7, and 11. Use the single complete method to classify them.

Solution

Initially, each sample forms its own cluster. We use Euclidean distance and follow these steps:

1. Calculate the distances between the 5 clusters to obtain the initial inter-cluster distance matrix $D^{(0)}$.

	G_1	G_2	G_3	G_4	G_5
G_1	0.0				
G_2	1.0	0.0			
G_3	5.5	4.5	0.0		
G_4	6.0	5.0	0.5	0.0	
G_5	10.0	9.0	4.5	4.0	0.0

Example 2

Step 2: Merge Closest Clusters

The smallest off-diagonal element in $D^{(0)}$ is $D_{34} = 0.5$. Therefore, merge G_3 and G_4 into a new cluster $G_6 = G_3 \cup G_4 = \{6.5, 7\}$. Calculate distances between G_6 and other clusters using the recursive formula for Complete Linkage. Obtain the new distance matrix $D^{(1)}$.

Distance Matrix $D^{(1)}$ (Complete Linkage)

	G_1	G_2	G_5	G_6
G_1	0.0			
G_2	1.0	0.0		
G_5	10.0	9.0	0.0	
G_6	6.0	5.0	4.5	0.0

Where: $G_1 = \{1\}, G_2 = \{2\}, G_5 = \{11\}, G_6 = \{6.5, 7\}$

Example 2

Step 3: Merge Next Closest Clusters

The smallest off-diagonal element in $D^{(1)}$ is $D_{12} = 1.0$. Therefore, merge G_1 and G_2 into a new cluster $G_7 = G_1 \cup G_2 = \{1, 2\}$. Calculate distances between G_7 and other clusters using the recursive formula for Complete Linkage. Obtain the new distance matrix $D^{(2)}$.

Distance Matrix $D^{(2)}$ (Complete Linkage)

	G_5	G_6	G_7
G_5	0.0		
G_6	4.5	0.0	
G_7	10.0	6.0	0.0

Where: $G_5 = \{11\}, G_6 = \{6.5, 7\}, G_7 = \{1, 2\}$

Step 4: Merge Next Closest Clusters

The smallest off-diagonal element in $D^{(2)}$ is $D_{56} = 4.5$. Therefore, merge G_5 and G_6 into a new cluster $G_8 = G_6 \cup G_5 = \{6.5, 7, 11\}$. Calculate distances between G_8 and other clusters using the recursive formula (5.6) for Complete Linkage. Obtain the new distance matrix $D^{(3)}$

Distance Matrix $D^{(3)}$ (Complete Linkage)

	G_7	G_8
G_7	0.0	
G_8	10.0	0.0

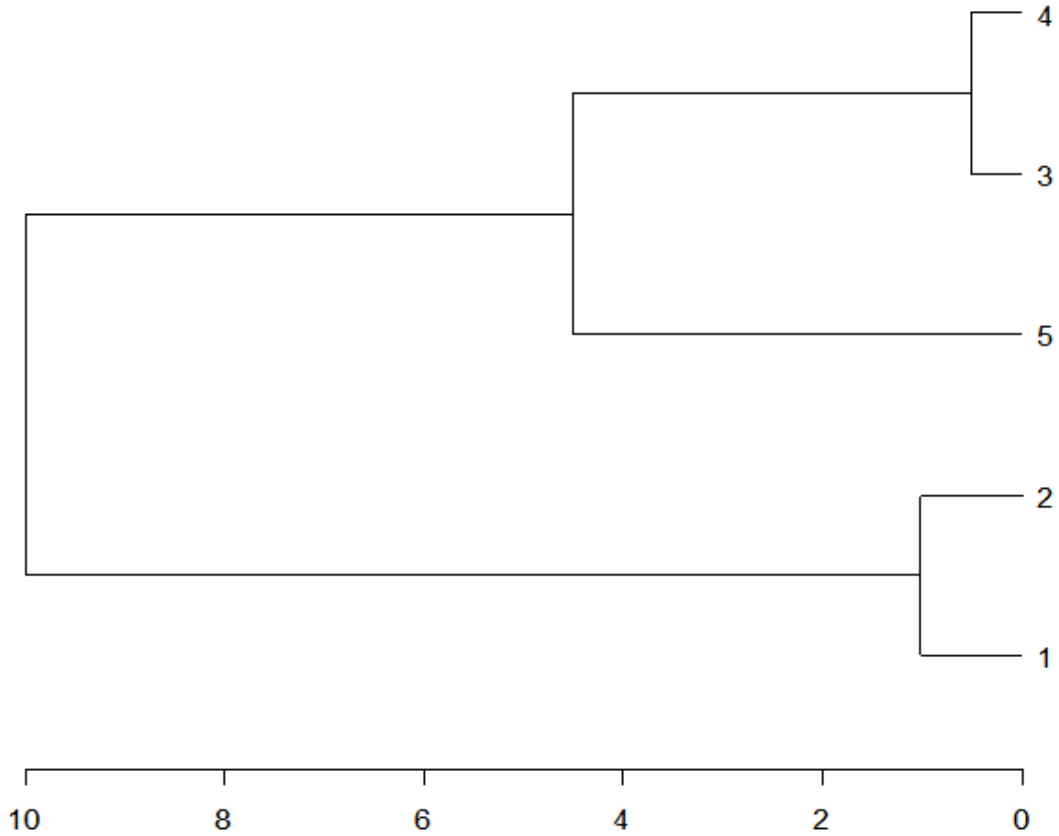
Where: $G_7 = \{1, 2\}$, $G_8 = \{6.5, 7, 11\}$

Step 5: Final Merge

Finally, merge G_7 and G_8 into G_9, at which point all 5 samples are merged into one cluster, and the clustering process terminates.

Example 2

- **If divided into 2 clusters:**
  - Cluster 1: $G_1 \cup G_2 = \{1, 2\}$
  - Cluster 2: $G_3 \cup G_4 \cup G_5 = \{6.5, 7, 11\}$
- **If divided into 3 clusters:**
  - Cluster 1: $G_1 \cup G_2 = \{1, 2\}$
  - Cluster 2: $G_5 = \{11\}$
  - Cluster 3: $G_3 \cup G_4 = \{6.5, 7\}$



Median Method (Intermediate Distance)

Definition

The distance between two clusters is defined as an intermediate value between the distance of their closest elements and the distance of their farthest elements. This approach is called the **Median Method**.

Recursive Formula

When clusters G_p and G_q merge to form a new cluster G_r, the squared distance between G_r and any other cluster G_k can be computed using the recursive formula:

$$D_{rk}^2 = \frac{1}{2}(D_{pk}^2 + D_{qk}^2) - \frac{1}{4}D_{pq}^2$$

Average Linkage Method

Definition

The Average Linkage Method defines the distance between two clusters as the average of all pairwise distances between their elements. Specifically, the distance between clusters G_p and G_q is defined as:

$$D_{pq} = \frac{1}{n_p n_q} \sum_{i \in G_p, j \in G_q} d_{ij},$$

where n_p and n_q are the number of samples in clusters G_p and G_q.

Recursive Formula

When clusters G_p and G_q merge to form a new cluster G_r, the distance between G_r and any other cluster G_k can be computed using the recursive formula:

$$D_{rk} = \frac{n_p}{n_p + n_q} D_{pk} + \frac{n_q}{n_p + n_q} D_{qk}$$

Centroid Method

Definition

Let the centroids of clusters G_p and G_q be $\bar{\boldsymbol{x}}_p$ and $\bar{\boldsymbol{x}}_q$, then the squared distance between G_p and G_q is defined as:

$$D_{pq}^2 = d_{\bar{\boldsymbol{x}}_p \bar{\boldsymbol{x}}_q}^2 = (\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_q)^t (\bar{\boldsymbol{x}}_p - \bar{\boldsymbol{x}}_q).$$

This hierarchical clustering method is called the **Centroid Method**.

Recursive Formula

G_p and G_q merge to form a new cluster G_r, the squared distance between G_r and any other cluster G_k can be computed using the recursive formula:

$$D_{rk}^2 = \frac{n_p}{n_r} D_{pk}^2 + \frac{n_q}{n_r} D_{qk}^2 - \frac{n_p n_q}{n_r^2} D_{pq}^2$$

where n_p, n_q, and n_r are the sizes of clusters G_p, G_q, and G_r respectively, with $n_r = n_p + n_q$.

Ward's Method (Minimum Variance Method)

Introduction

Ward's Method, is based on the idea of analysis of variance: If clusters are correctly formed, the within-cluster sum of squares should be small.

Distance Definition

The squared distance between clusters G_p and G_q is defined as:

$$D_{pq}^2 = W_r - W_p - W_q$$

where

$$W_p = \sum_{i \in G_p} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_p)^t (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_p), \quad W_q = \sum_{i \in G_q} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_q)^t (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_q)$$

$$W_r = \sum_{i \in G_r} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_r)^t (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_r), \quad \text{with} G_r = G_p \cup G_q.$$

Ward's Method (Continued)

Recursive Formula

When clusters G_p and G_q merge to form a new cluster G_r, the squared distance between G_r and any other cluster G_k can be computed using the recursive formula:

$$D_{rk}^2 = \frac{(n_k + n_p)}{n_k + n_r} D_{kp}^2 + \frac{(n_k + n_q)}{n_k + n_r} D_{kq}^2 - \frac{n_k}{n_k + n_r} D_{pq}^2$$

where n_p, n_q, n_r, and n_k are the sizes of the respective clusters.

Recursive Formulas

Method	Recursive Formula
(1) Single Linkage	$D_{kr} = \min\{D_{kp}, D_{kq}\}$
(2) Complete Linkage	$D_{kr} = \max\{D_{kp}, D_{kq}\}$
(3) Median Method	$D_{kr}^2 = \frac{1}{2}D_{kp}^2 + \frac{1}{2}D_{kq}^2 - \frac{1}{4}D_{pq}^2$
(4) Average Linkage	$D_{kr}^2 = \frac{n_p}{n_r}D_{kp}^2 + \frac{n_q}{n_r}D_{kq}^2$
(5) Centroid Method	$D_{kr}^2 = \frac{n_p}{n_r}D_{kp}^2 + \frac{n_q}{n_r}D_{kq}^2 - \frac{n_p n_q}{n_r^2}D_{pq}^2$
(6) Ward's Method	$D_{kr}^2 = \frac{n_k+n_p}{n_k+n_r}D_{kp}^2 + \frac{n_k+n_q}{n_k+n_r}D_{kq}^2 - \frac{n_k}{n_k+n_r}D_{pq}^2$

Notation:

- G_p and G_q merge to form G_r ($G_r = G_p \cup G_q$)
- n_p, n_q, n_r, n_k: number of elements in clusters G_p, G_q, G_r, G_k
- D_{ij}: distance between clusters G_i and G_j

Key Characteristics:

- **Single:** Uses minimum distance (nearest neighbor)
- **Complete:** Uses maximum distance (farthest neighbor)
- **Median:** Intermediate distance, compromise approach
- **Average:** Uses average of all pairwise distances
- **Centroid:** Uses distance between cluster centers
- **Ward:** Minimizes within-cluster variance increase

Example 3: Regional Consumption Pattern Analysis

Objective

Find the distribution patterns of urban household consumption expenditure across 22 provinces, municipalities, and autonomous regions in China in 2018, and classify these regions into consumption types using cluster analysis.

Data Description

The dataset includes annual per capita consumption expenditure for urban households across 8 major categories:
(1) Food, alcohol, and tobacco **(2)** Clothing **(3)** Housing **(4)** Household goods and services **(5)** Transportation and communication **(6)** Education, culture, and entertainment **(7)** Health care and medical services **(8)** Other goods and services.

Urban Household Per Capita Consumption Expenditure Data (2018) (Unit: Yuan per person)

Region	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Beijing	8064.9	2175.5	14110.3	2371.9	4767.4	3999.4	3274.5	1078.6
Tianjin	8647.5	1990.0	6406.3	1818.4	4280.9	3186.6	2676.9	896.3
Hebei	4271.3	1257.4	4050.4	1138.7	2355.4	1734.5	1540.5	373.8
Shanxi	3688.2	1261.0	3228.5	855.6	1845.2	1940.0	1635.1	356.4
Inner Mongolia	5324.3	1751.2	3680.0	1204.6	3074.3	2245.4	1847.5	537.9

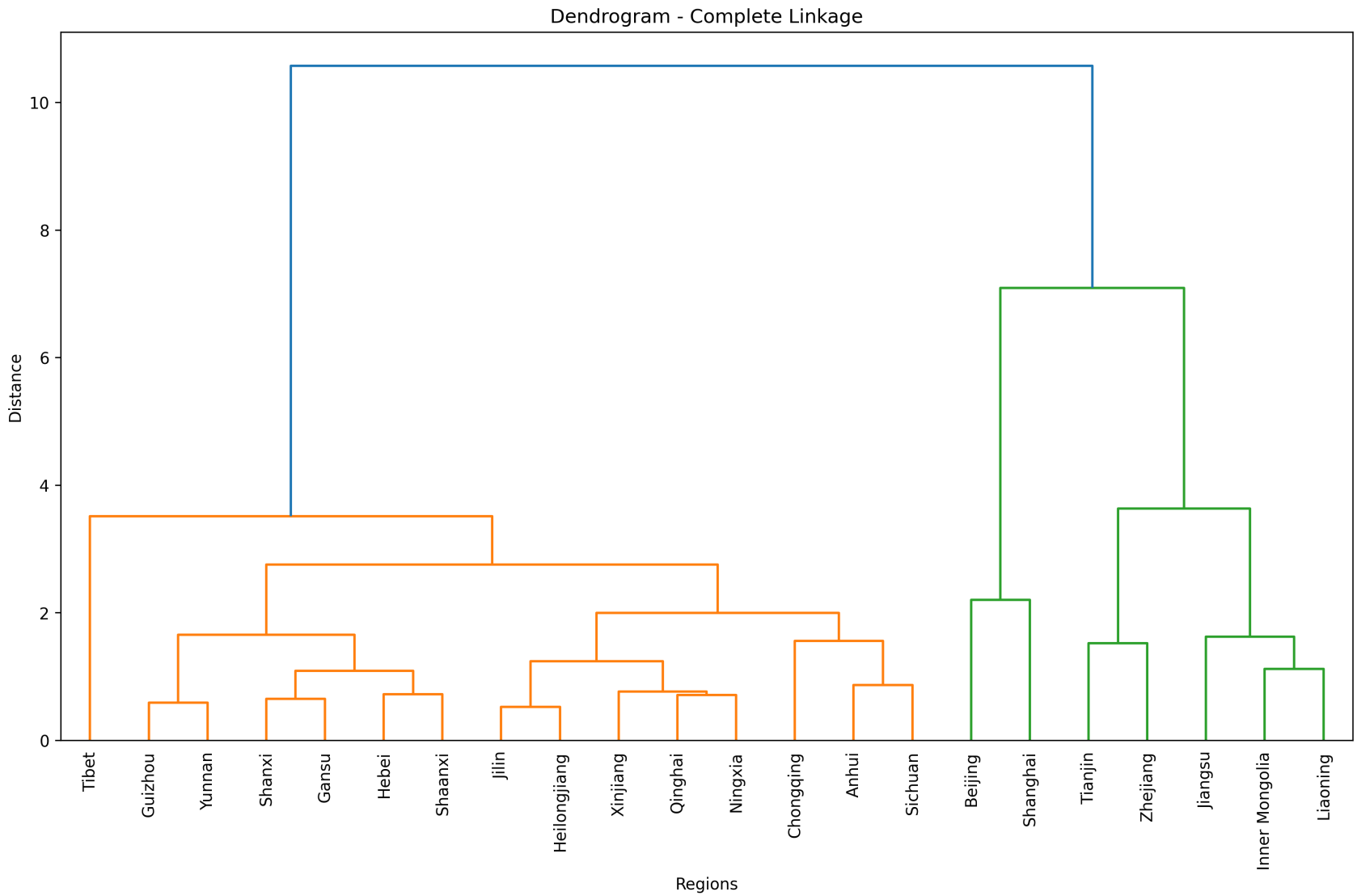
Analysis Method

- Hierarchical clustering will be applied to classify the 22 regions
- The 8 consumption categories serve as clustering variables
- Different linkage methods can be compared for optimal clustering
- Results will help identify regional consumption patterns and economic development levels

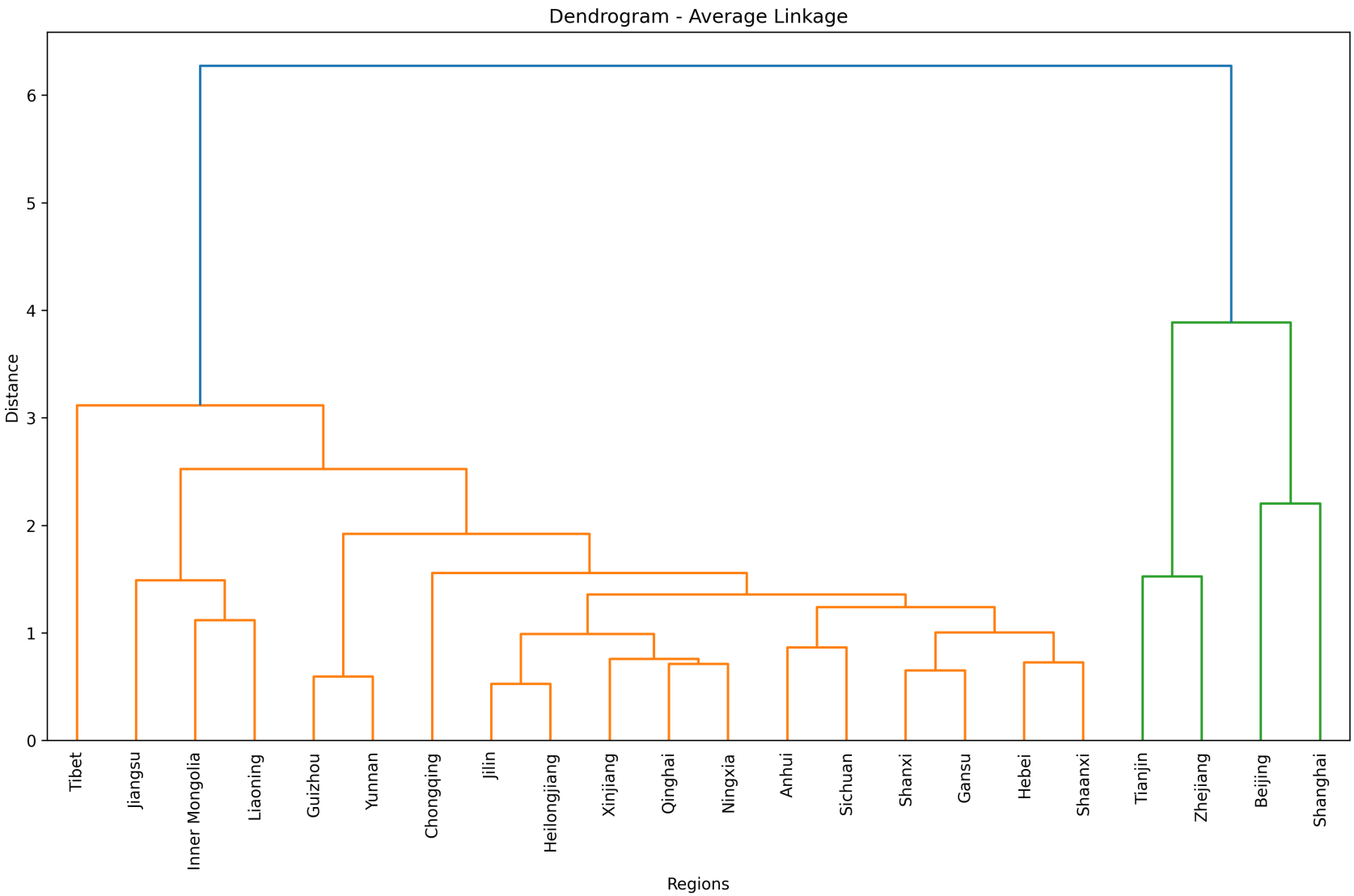
Urban Household Per Capita Consumption Expenditure Data (2018) - Continued

Region	Food, Alcohol Tobacco	Clothing	Housing	Household Goods & Services	Transportation & Communication	Education Culture & Entertainment	Health Care	Other Goods & Services
Liaoning	5727.8	1628.1	4169.5	1259.4	2968.2	2708.0	2257.1	680.2
Jilin	4417.4	1397.0	3294.8	899.4	2479.7	2193.4	2012.0	506.7
Heilongjiang	4573.2	1405.4	3176.3	866.4	2196.6	2030.3	2235.3	490.4
Shanghai	10728.2	2036.8	14208.5	2095.5	4881.2	5049.4	3070.2	1281.5
Jiangsu	6529.8	1541.0	6731.2	1493.3	3522.8	2582.6	2016.4	590.4
Zhejiang	8198.3	1813.5	7721.2	1652.4	4302.0	3031.3	2059.4	692.6
Anhui	5414.7	1137.4	3941.9	1041.2	2082.1	1810.4	1224.0	392.8
Chongqing	6220.8	1454.5	3498.8	1338.9	2545.0	2087.8	1660.0	442.8
Sichuan	5937.9	1173.8	3368.0	1182.2	2398.8	1599.7	1568.6	434.5
Guizhou	3792.9	934.7	2760.7	878.1	2408.0	1660.0	1083.5	280.1
Yunnan	3983.4	789.1	3081.1	859.9	2212.8	1772.7	1267.7	283.2
Tibet	4330.5	1285.2	2102.6	622.3	1847.7	609.3	460.1	262.6
Shaanxi	4292.5	1141.1	3388.2	1200.8	2005.8	2008.8	1749.4	373.2
Gansu	4253.3	1111.5	3095.0	896.9	1640.7	1710.3	1573.9	342.4
Qinghai	4671.6	1350.6	2990.0	932.0	2671.4	1655.6	1842.0	444.0
Ningxia	4234.1	1388.2	3014.3	1067.1	2724.4	2139.5	1727.1	420.4
Xinjiang	4691.6	1456.0	2894.3	1082.8	2274.4	1762.5	1592.6	434.9

Complete Linkage



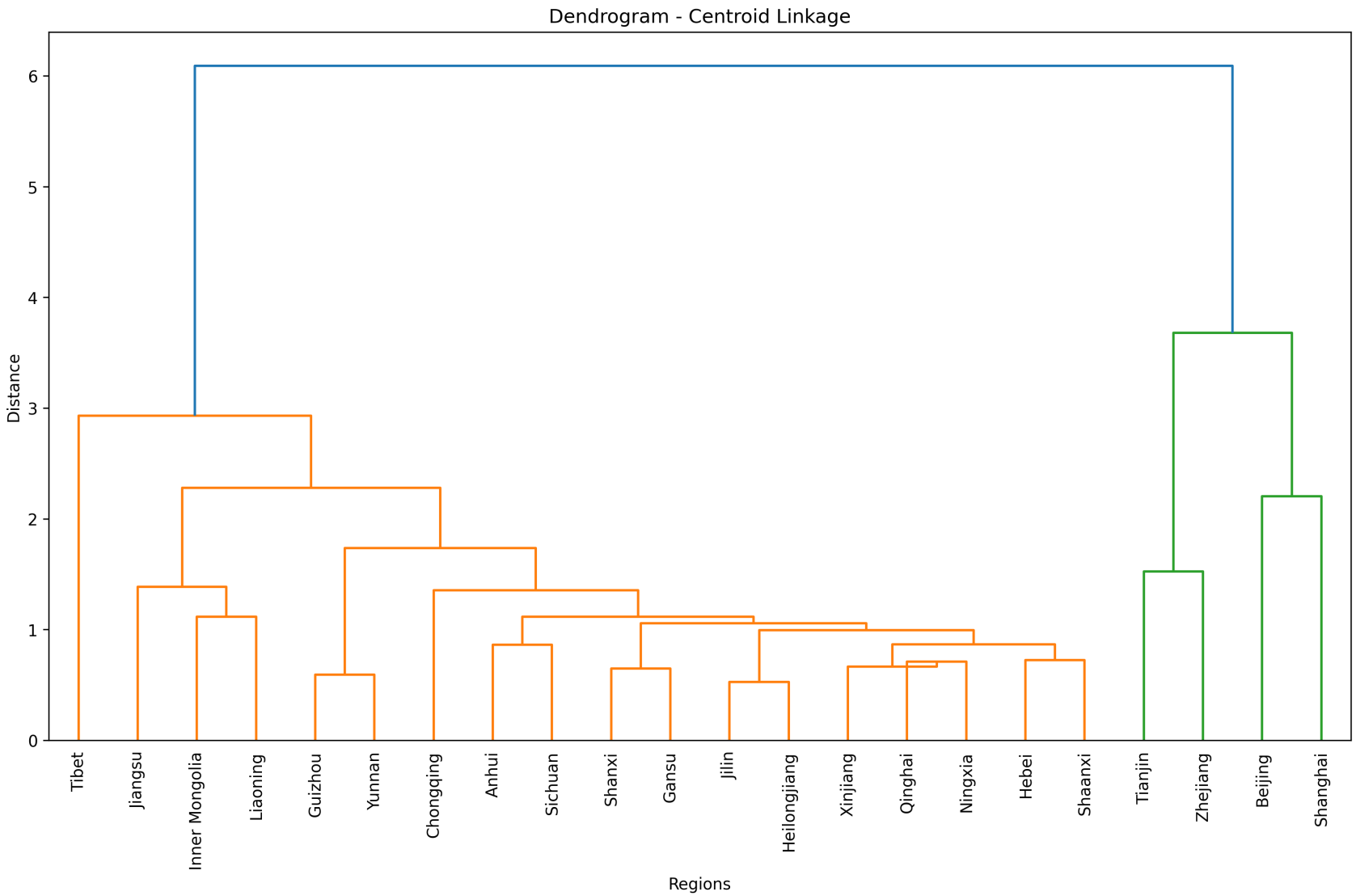
Average Linkage



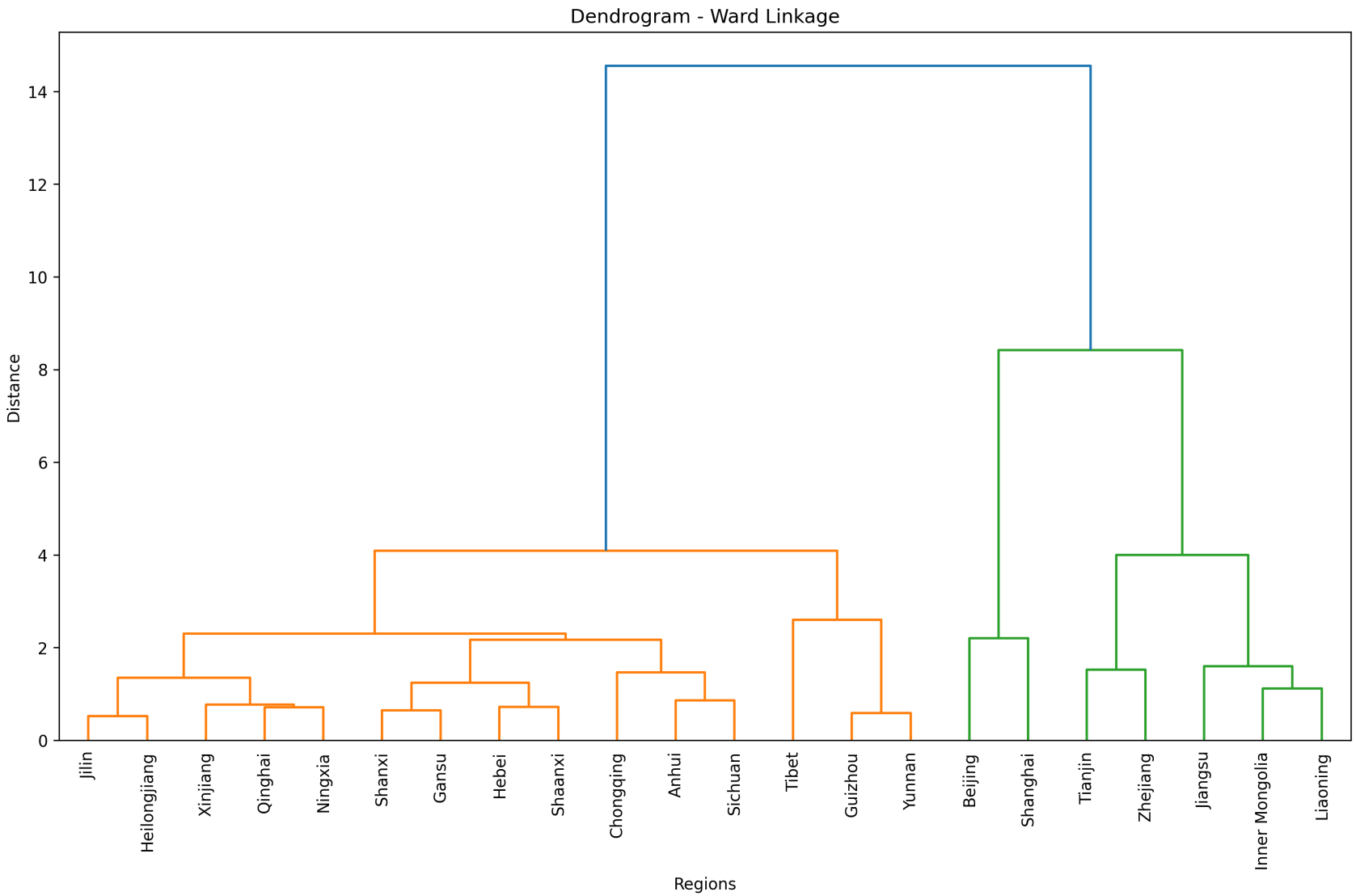
Example 3

Based on the clustering results of the **Complete Linkage method** and the **Average Linkage method**, the 22 regions can be roughly divided into three categories: **Beijing, Shanghai, Tianjin, and Zhejiang** are grouped in the first category, which are high-consumption areas; **Jiangsu, Inner Mongolia, and Liaoning** are grouped in the second category, which are second-highest-consumption areas; and the **remaining provinces** are grouped in the first category, which are low- to medium-consumption areas. **Tibet** is a special case and could also be classified as a separate category, or not be categorized for the time being.

Centroid


Dendrogram - Centroid Linkage
Distance
0
1
2
3
4
5
6
Tibet
Jiangsu
Inner Mongolia
Liaoning
Guizhou
Yunnan
Chongqing
Anhui
Sichuan
Shanxi
Gansu
Jilin
Heilongjiang
Xinjiang
Qinghai
Ningxia
Hebei
Shaanxi
Tianjin
Zhejiang
Beijing
Shanghai
Regions

Ward


Dendrogram - Ward Linkage
Distance
0
2
4
6
8
10
12
14
Jilin
Heilongjiang
Xinjiang
Qinghai
Ningxia
Shanxi
Gansu
Hebei
Shaanxi
Chongqing
Anhui
Sichuan
Tibet
Guizhou
Yunnan
Beijing
Shanghai
Tianjin
Zhejiang
Jiangsu
Inner Mongolia
Liaoning
Regions

Example 3

Comparing the clustering results of the **centroid method** and the **Ward method** reveals that the centroid method results show a significant difference in the number of samples in each category, while the Ward method's clustering results are more balanced, thus favoring the Ward method.

According to the Ward method, the 22 regions can be roughly divided into four categories: **Beijing, Shanghai, Tianjin, and Zhejiang** are grouped in the first category, representing high-consumption areas; **Jiangsu, Inner Mongolia, and Liaoning** are grouped in the second category, representing the second-highest-consumption areas; **Hubei, Chongqing, Shanxi, Henan, Gansu, Sichuan, Hebei, and Shaanxi** are grouped in the third category, representing medium-consumption areas; and all **other regions** are grouped in the fourth category, representing low-consumption areas.

Dynamic Clustering Method

Basic Concept

The dynamic clustering method is also called the stepwise clustering method. Its basic idea is to first select a group of clustering points or give an initial classification, let the samples cluster to the clustering points according to a certain principle, and continuously update or iterate the clustering points until the classification is more reasonable or the iteration reaches stability, thus forming a final classification result.

K-means Clustering Method

There are many different methods for dynamic clustering, the most popular of which is k-means clustering. It is a fast clustering algorithm proposed and named by MacQueen (1967). It uses a partitioning approach to achieve clustering.

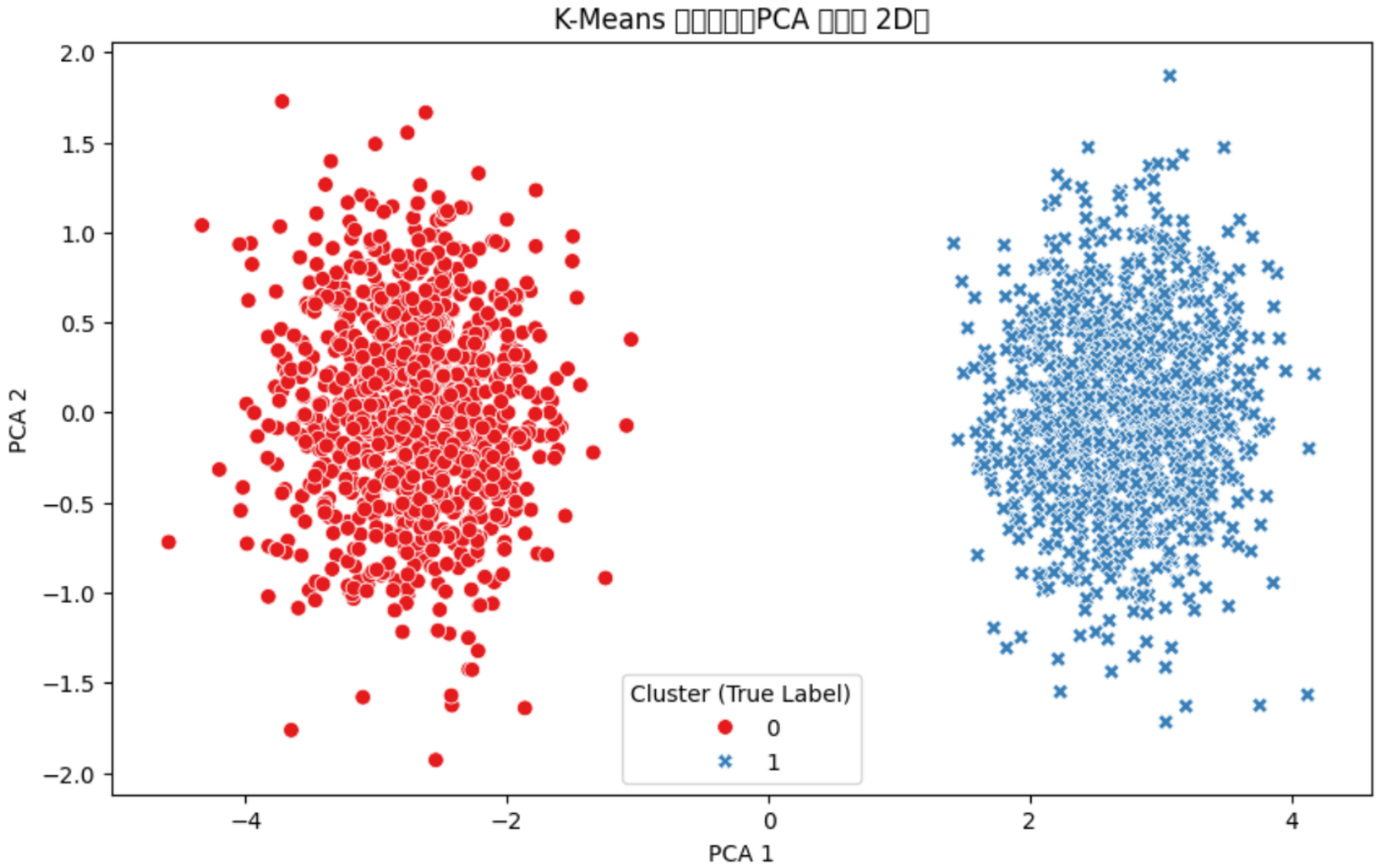
K-means Clustering Algorithm: Step-by-Step Procedure

1. **Specify the number of clusters (k)**
   - Determine the appropriate number of clusters for the analysis
2. **Determine k initial centers**
   - Select k initial cluster centers or divide all samples into k initial clusters and then use the centroids (means) of these initial clusters as starting points
3. **Assign samples to clusters based on proximity**
   - Calculate distances from each observation to all k centroids
   - Assign each sample to the nearest cluster using the closest centroid principle to form k clusters.
4. **Update cluster centroids**
   - Recalculate the centroid for each cluster
   - Compute the mean of all sample points within each cluster
   - Use these new centroids as updated cluster centers
5. **Check convergence criteria**
   - Determine if termination conditions are satisfied
   - If not converged, return to Step 3 and repeat the process
   - Continue iterations until convergence criteria are met

Example 4

Simulation analysis using the kmeans algorithm in Python. To demonstrate the advantages of the kmeans method for large sample data, we first simulate 10,000 normally distributed random data with a mean of 0 and a standard deviation of 0.3. We then convert these data into a matrix with 10 variables and 1,000 samples. Using the same method, we simulate 10,000 normally distributed random data with a mean of 1 and a standard deviation of 0.3. We then convert these data into a matrix with 10 variables and 1,000 samples. Finally, we merge these two matrices into a matrix with 10 variables and 2,000 samples. We then use the kmeans clustering method to cluster the data into two clusters and observe the clustering effect.

K-Means for Example 4



K-Means for Example 3

