Module 4

Models for Multivariate Ordinal Data

MESIO Summer School

Introduction to Multivariate Ordinal Data

Limitations of Classical Multivariate Methods

Techniques for Dimensionality Reduction

Clustering with Ordinal and Mixed Data

Recap: What is Ordinal Data?

- We've previously discussed **ordinal data**: categorical variables with a meaningful order but unequal or unknown intervals between categories.
- Examples: Likert scales (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree), educational levels (Primary, Secondary, University).
- Previous models: **Proportional Odds Models**, **CUB models** for single or few ordinal variables, often for explanation/prediction.

The Need for Multivariate Approaches

- Many real-world scenarios involve **multiple ordinal variables** collected simultaneously.
- **Examples:**
  - Customer satisfaction surveys (rating multiple services/products).
  - Psychological questionnaires (measuring various aspects of well-being/attitude).
- **Goal:** Not just modeling a single response, but uncovering patterns, similarities, or latent structures across multiple variables.

Fundamental Goals in Multivariate Data Analysis

Before specific methods, let's define two key goals:

1. **Dimensionality Reduction**:
   - Transforming high-dimensional data into a lower-dimensional space.
   - Aims to simplify, visualize, and interpret data by retaining essential information.
   - Often uncovers underlying "latent" variables or themes.

Fundamental Goals in Multivariate Data Analysis

Before specific methods, let’s define two key goals:

1. **Dimensionality Reduction**:
   - Transforming high-dimensional data into a lower-dimensional space.
   - Aims to simplify, visualize, and interpret data by retaining essential information.
   - Often uncovers underlying ”latent” variables or themes.
2. **Clustering**:
   - An unsupervised learning technique.
   - Finding natural groupings (clusters) within a dataset.
   - Observations within a group are more similar to each other than to those in other groups.

Classical Methods: Limitations for Ordinal Data

- Classical multivariate techniques like **Principal Component Analysis (PCA)** and **K-means clustering** are powerful for *continuous data*.
- **Challenge:** Ordinal scales represent ordered categories, not true measurable quantities with uniform intervals.
- **Important:** Respecting this fundamental distinction to avoid misleading conclusions.

Introduction to Multivariate Ordinal Data

Limitations of Classical Multivariate Methods

Techniques for Dimensionality Reduction

Clustering with Ordinal and Mixed Data

PCA and Ordinal Data

- **PCA:** Dimensionality reduction based on covariance/correlation matrix.
- Assumes data are numerical and distances between values are meaningful and consistent.
- **Problem with Ordinal Data:**
  - Treats numerical labels (e.g., 1, 2, 3, 4, 5) as continuous, interval-scaled quantities.
  - Example: Distance between ”Agree” (4) and ”Strongly Agree” (5) is treated as identical to ”Neutral” (3) and ”Agree” (4).
  - These ”distances” on an ordinal scale are rarely equal in true conceptual magnitude.
- **Consequence:** PCA can generate principal components that distort the true underlying structure, leading to misleading interpretations of latent dimensions.

K-Means Clustering and Ordinal Data

- **Standard K-means:** Relies on **Euclidean distances** to quantify similarity.
- Euclidean distance assumes interval-scaled variables where differences are direct and comparable.
- **Problem with Ordinal Data:**
  - Calculated Euclidean distances are based on arbitrary numerical assignments.
  - They do not reflect the true, often unequal, conceptual distances between ordered categories.
- **Consequence:**
  - Formation of artificial clusters that do not genuinely reflect meaningful patterns.
  - Relevant patterns might be masked, or spurious groupings might emerge.
  - **Conclusion:** For multivariate ordinal data, analytical tools must account for their ordered nature.

Introduction to Multivariate Ordinal Data

Limitations of Classical Multivariate Methods

Techniques for Dimensionality Reduction

Clustering with Ordinal and Mixed Data

Overview of Ordinal-Friendly Dimensionality Reduction

- These methods aim to reduce dimensionality and provide insightful graphical representations.
- They uncover latent structures and visualize patterns in a lower-dimensional space.
- We will discuss:
  - **Nonlinear Principal Component Analysis (NLPCA)**
  - **Correspondence Analysis (CA)**
  - **Multiple Correspondence Analysis (MCA)**

Nonlinear Principal Component Analysis (NLPCA)

- **Advancement over classical PCA**: Extends its framework for diverse data types, including ordinal.
- **Core Idea:** Transforms original categorical responses into **optimal scores**.
- **How it works:**
  - Iteratively determines the most appropriate numerical values (optimal scores) for each category.
  - These scores maximize the variance explained by the resulting principal components, while preserving category order.
  - Captures potentially nonlinear relationships.
- **Benefits:**
  - Effectively handles mixed data (continuous, nominal, ordinal).
  - Output reflects major sources of variability, respecting intrinsic ordinal nature.
  - Valuable when latent constructs exhibit nonlinear relationships.

Correspondence Analysis (CA)

- **Purpose:** Explores association between **two categorical variables** (contingency table).
- **Strength:** Graphically represents relationships between row and column categories in a common low-dimensional space (e.g., 2D).
- Proximity on the map reflects association (tendency to co-occur).
- **Example:** ”Education Level” (ordinal) vs. ”Preferred News Source” (nominal/ordinal). CA reveals if specific education levels are associated with particular news sources.

Correspondence Analysis (CA)

- **Purpose:** Explores association between **two categorical variables** (contingency table).
- **Strength:** Graphically represents relationships between row and column categories in a common low-dimensional space (e.g., 2D).
- Proximity on the map reflects association (tendency to co-occur).
- **Example:** ”Education Level” (ordinal) vs. ”Preferred News Source” (nominal/ordinal). CA reveals if specific education levels are associated with particular news sources.
- **Limitation for Ordinal Data:** Standard CA does *not explicitly leverage or enforce the order information* of ordinal variables.
- Still a valuable **exploratory tool** for general patterns and broad relationships, especially in initial data exploration.

Multiple Correspondence Analysis (MCA)

- **Extension of CA:** Analyzes relationships among more than two categorical variables simultaneously.
- Facilitates detection of latent structures and similarities:
  - Identifies clusters of individuals with similar response profiles.
  - Reveals underlying dimensions or ”themes” explaining relationships among survey items.
- Widely used in social sciences, marketing, psychology for dimensionality reduction and visualization.

Multiple Correspondence Analysis (MCA)

- **Extension of CA:** Analyzes relationships among more than two categorical variables simultaneously.
- Facilitates detection of latent structures and similarities:
  - Identifies clusters of individuals with similar response profiles.
  - Reveals underlying dimensions or ”themes” explaining relationships among survey items.
- Widely used in social sciences, marketing, psychology for dimensionality reduction and visualization.
- **Key Point:** Standard MCA traditionally treats all variables as **nominal**, disregarding inherent order.

Multiple Correspondence Analysis (MCA)

- **Extension of CA:** Analyzes relationships among more than two categorical variables simultaneously.
- Facilitates detection of latent structures and similarities:
  - Identifies clusters of individuals with similar response profiles.
  - Reveals underlying dimensions or ”themes” explaining relationships among survey items.
- Widely used in social sciences, marketing, psychology for dimensionality reduction and visualization.
- **Key Point:** Standard MCA traditionally treats all variables as **nominal**, disregarding inherent order.
- **Extensions (e.g., ordinal MCA, nonlinear MCA):** Incorporate specific coding/weighting schemes to account for ordering, producing more meaningful dimensions faithful to the data’s structure.

Introduction to Multivariate Ordinal Data

Limitations of Classical Multivariate Methods

Techniques for Dimensionality Reduction

Clustering with Ordinal and Mixed Data

Challenges for Clustering Ordinal Data

- As discussed, standard algorithms (e.g., K-means) are poorly suited for ordinal or mixed-type data.
- This is due to their reliance on distance metrics that assume continuous, interval-scaled variables.
- To overcome this, we need appropriate distance measures and clustering algorithms.

Distance-Based Clustering with Ordinal Data

- **Critical First Step:** Define distance measures that genuinely respect the ordinal nature.

Distance-Based Clustering with Ordinal Data

- **Critical First Step:** Define distance measures that genuinely respect the ordinal nature.
- **Gower's Distance:**
  - Widely adopted and versatile.
  - Handles **mixed data types** (nominal, ordinal, continuous) within a single dissimilarity metric.
  - For ordinal variables, it calculates distances preserving order and accounting for number of categories, without assuming equal intervals.

Distance-Based Clustering with Ordinal Data

- **Critical First Step:** Define distance measures that genuinely respect the ordinal nature.
- **Gower's Distance:**
  - Widely adopted and versatile.
  - Handles **mixed data types** (nominal, ordinal, continuous) within a single dissimilarity metric.
  - For ordinal variables, it calculates distances preserving order and accounting for number of categories, without assuming equal intervals.
- **Other Ordinal-Specific Measures:** Based on ranks, monotonic transformations, or agreement/disagreement counts.

Distance-Based Clustering with Ordinal Data

- **Critical First Step:** Define distance measures that genuinely respect the ordinal nature.
- **Gower's Distance:**
  - Widely adopted and versatile.
  - Handles **mixed data types** (nominal, ordinal, continuous) within a single dissimilarity metric.
  - For ordinal variables, it calculates distances preserving order and accounting for number of categories, without assuming equal intervals.
- **Other Ordinal-Specific Measures:** Based on ranks, monotonic transformations, or agreement/disagreement counts.
- **Clustering Algorithms (after computing dissimilarity matrix):**
  - **Hierarchical Clustering:** Builds a hierarchy of clusters based on dissimilarities.
  - **Partitioning Around Medoids (PAM):** Robust alternative to K-means; uses actual data points (medoids) as cluster centers, minimizing sum of dissimilarities (not squared Euclidean distances).

Model-Based Clustering for Ordinal Data

- **Assumption:** Data points within each cluster are generated from a specific probability distribution.
- **Goal:** Estimate parameters of component distributions and assign observations probabilistically.
- **For Ordinal Data:** Employs specialized distributions for ordered categories.
- **Example: Latent Class Models** (LCMs)
  - Each latent class represents a cluster.
  - Within each cluster, ordinal responses follow a discrete probability distribution.
- **Advantages:**
  - Provides a **statistical basis** for clustering (formal tests, information criteria like BIC/AIC for optimal number of clusters).
  - Handles **uncertainty in cluster membership** (probabilistic assignments).

Key Takeaways from the Course

This course provided an introduction to the analysis of ordinal data, crucial in social sciences, psychology, marketing, and applied domains.
We covered:

- **Definition and Characteristics of Ordinal Data:** Emphasizing the importance of appropriate treatment (not assuming interval/continuous scale properties).
- **Survey and Scale Design:** (Implicitly covered through the discussion of ordinal data properties).
- **Statistical Models for Ordinal Data:** Proportional Odds Model and CUB models.
- **Multivariate Methods for Ordinal and Mixed Data:** Dimensionality reduction and clustering techniques specifically designed or adapted for ordinal data.

Final Message

- A common thread throughout the course: **the need to respect the nature of ordinal data.**
- Choosing analytical methods that reflect their structure and meaning is paramount.
- Applying models designed for other data types (e.g., continuous) can lead to **invalid interpretations**.
- **Appropriate methods** provide valuable insights into individual preferences, attitudes, and behaviors.

Thank You

Thank You!