Dataset Selection Guidelines

For Project 1: Statistical Estimation through SVD Analysis

Overview

Choosing an appropriate dataset is crucial for meaningful PCA analysis. Your dataset should exhibit properties that make PCA analysis informative and allow you to demonstrate understanding of the theoretical concepts.

Dataset Requirements

1. Size and Dimensionality

- Minimum 100 samples (rows)
- At least 10 features (columns)
- Features should be numeric
- No more than 30% missing values

2. Statistical Properties

- Features should have different scales and variances
- Mix of correlated and uncorrelated features
- No perfect collinearity between features
- Reasonable spread of variance across features

3. Dataset Options

Option 1: Provided Datasets

1. Climate Data
   - Daily weather measurements
   - Features: temperature, pressure, humidity, wind speed, etc.
   - Good for: demonstrating scale effects
2. Financial Market Data
   - Daily stock returns

- Features: different company returns
- Good for: correlation analysis

3. Sensor Measurements
   - IoT sensor readings
   - Features: multiple sensor types
   - Good for: noise analysis

Option 2: Your Own Dataset Must meet these criteria: - Satisfies size requirements - Contains numeric features - Has documentation/context - Approved by instructor

Analysis Considerations

1. Preprocessing Requirements

- Handle missing values appropriately
- Document outlier treatment
- Justify normalization choices

2. Context Importance

- Understand what features represent
- Know units and scales
- Consider domain-specific implications

3. Documentation Needs

- Source of data
- Any preprocessing steps
- Feature descriptions
- Known relationships

Submission Requirements

1. Dataset Proposal (Due with Milestone 1)

- Dataset description

- Basic statistics
- Preprocessing plan
- Expected challenges

2. Final Documentation

- Complete data description
- Preprocessing steps taken
- Justification for choices
- Limitations encountered

Tips for Success

- Start with exploratory data analysis
- Look for interesting feature relationships
- Consider interpretability
- Document assumptions