Statistical Analysis of Text in Survey Records

Wendy Martinez
Alex Measure

Bureau of Labor Statistics

FCSM
November 2013

BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

www.bls.gov

Preliminaries

- Acknowledge my Bureau of Labor Statistics (BLS) colleagues
- This work does not represent official BLS policy.
- Preliminary work – illustrate concepts

BLS

Outline

- Describe the objective and the data set
- Provide 'operational scenario'
- List steps for the analysis
- Discuss results of the analysis
- Offer some ideas for future work and applications

BLS

Objectives

- Explore ways we can analyze text in survey data
- **Examples**:
  - Computer assisted coding – Survey of Occupational Injuries and Illnesses (SOII)
  - Document clustering – Occupational Safety and Health reports
  - Document classification – Health benefits for National Compensation Survey (NCS)

BLS

Example – Accident Reports

- Downloaded accident reports from Department of Labor
  - Enforcement Data website
- Main site for data:
  - http://ogesdw.dol.gov/data_catalogs
- Click on OSHA download:
  - http://ogesdw.dol.gov/data_summary.php
- Data dictionary:
  - http://enforcedata.dol.gov/dd_display.php

BLS

Example – Accident Reports

- Files have the following:
  - Date of accident
  - Event description – short phrase
  - Event keywords – Nature, Part, Source, etc
  - Event type – label
  - Abstract – unstructured text
  - More …

BLS

Example – Accident Reports

- What can we do with these data?
- **Document clustering**:
  - Group abstracts – documents in group have similar topic
  - Use as an aid for coding, editing, verification
- **Document classification**:
  - Use labeled abstracts (e.g., event type) to build classifier for future abstracts *

BLS

* *This was not done for this presentation.*

Steps for Text Analysis

- **Step 1**: Preprocess Text
  - Remove special characters
  - Remove stop words
  - Stemming*
- **Step 2**: Encode Text
  - Term-document matrix (bag of words)
  - Bigram proximity matrix (BPMs) *
  - Term weighting *

BLS

* *This was not done for this presentation.*

Steps for Text Analysis

- **Step 3**: Reduce Dimensionality
  - Singular Value Decomposition (LSI/LSA) *
  - Isomap
  - Nonnegative Matrix Factorization *
- **Step 4**: Cluster Documents
  - Agglomerative Clustering – Hierarchical *
  - Model-Based Clustering – Finite Mixtures

BLS

* *This was not done for this presentation.*

Steps for Text Analysis

- **Step 5**: Assess Clusters
  - Visualization
  - Exploratory Data Analysis
  - Determine cluster topics

BLS

Step 1: Process the Data

- Extracted accident reports for November and December of 2011
- Preprocessed as described before
  - Removed special characters
  - Removed stop words
- Yielded 358 accident reports
- Lexicon had 3841 words

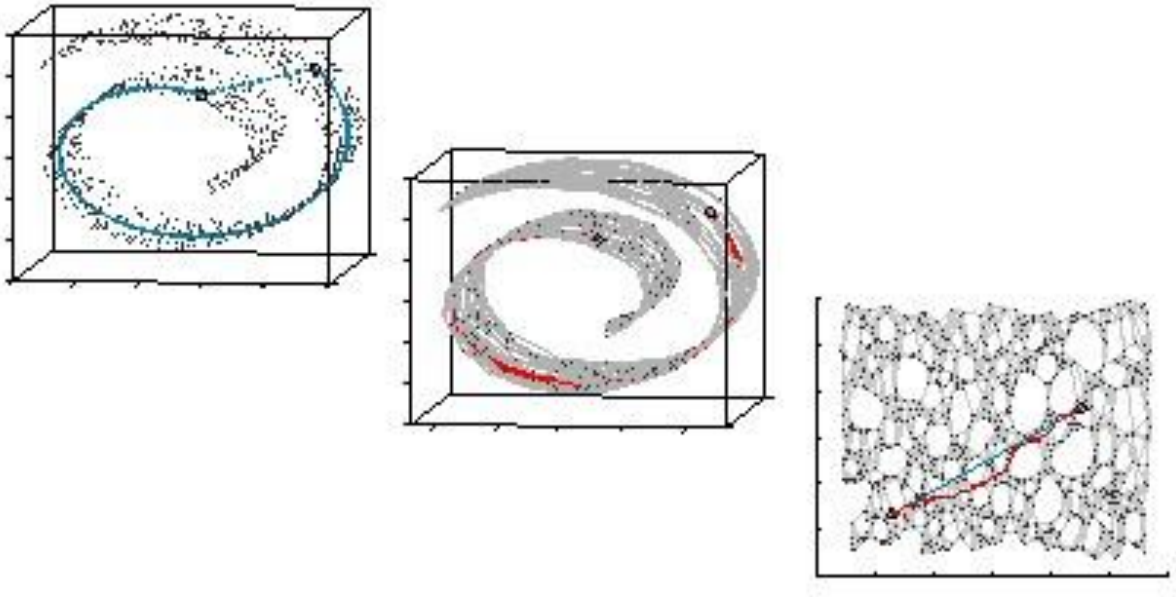
BLS

Step 2: Encode the Text

- The most common approach is the bag of words or term-document matrix.
- The rows correspond to words.
- The columns correspond to documents.
- The (i,j) -th entry in the matrix is the number of times the i -th word appears in the j -th document.

Step 3: Reduce Dimensions

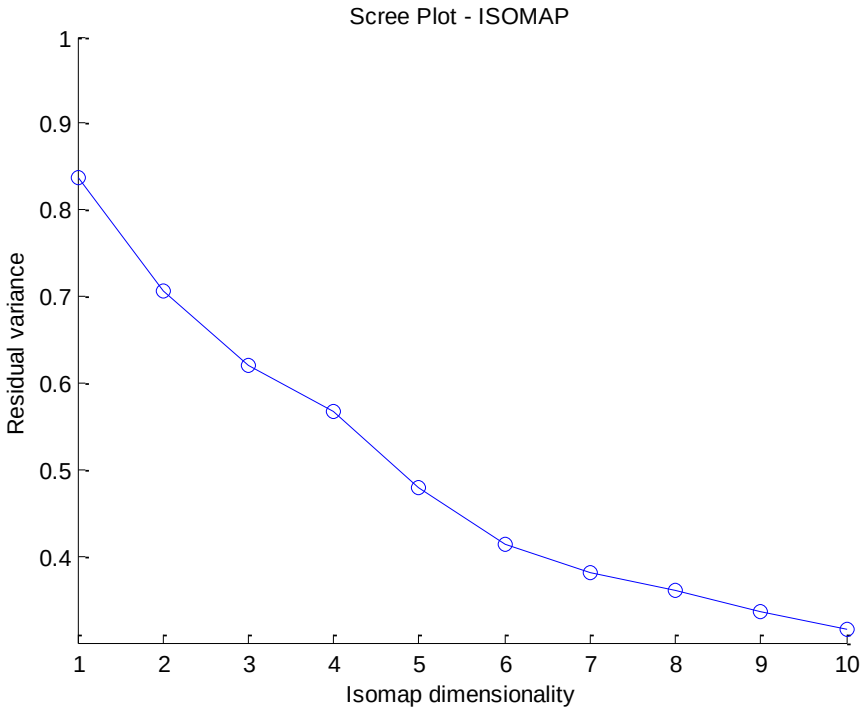
- Isomap: Nonlinear dimensionality reduction
- Classical multidimensional scaling
- Inputs are geodesic distances



BLS

Choosing the Number of Dimensions

- Use a scree plot
- Look for elbow in the curve
- Chose 4 dimensions
- Used EDA Toolbox for MATLAB to
  - Visualize
  - Cluster
  - Assess



BLS

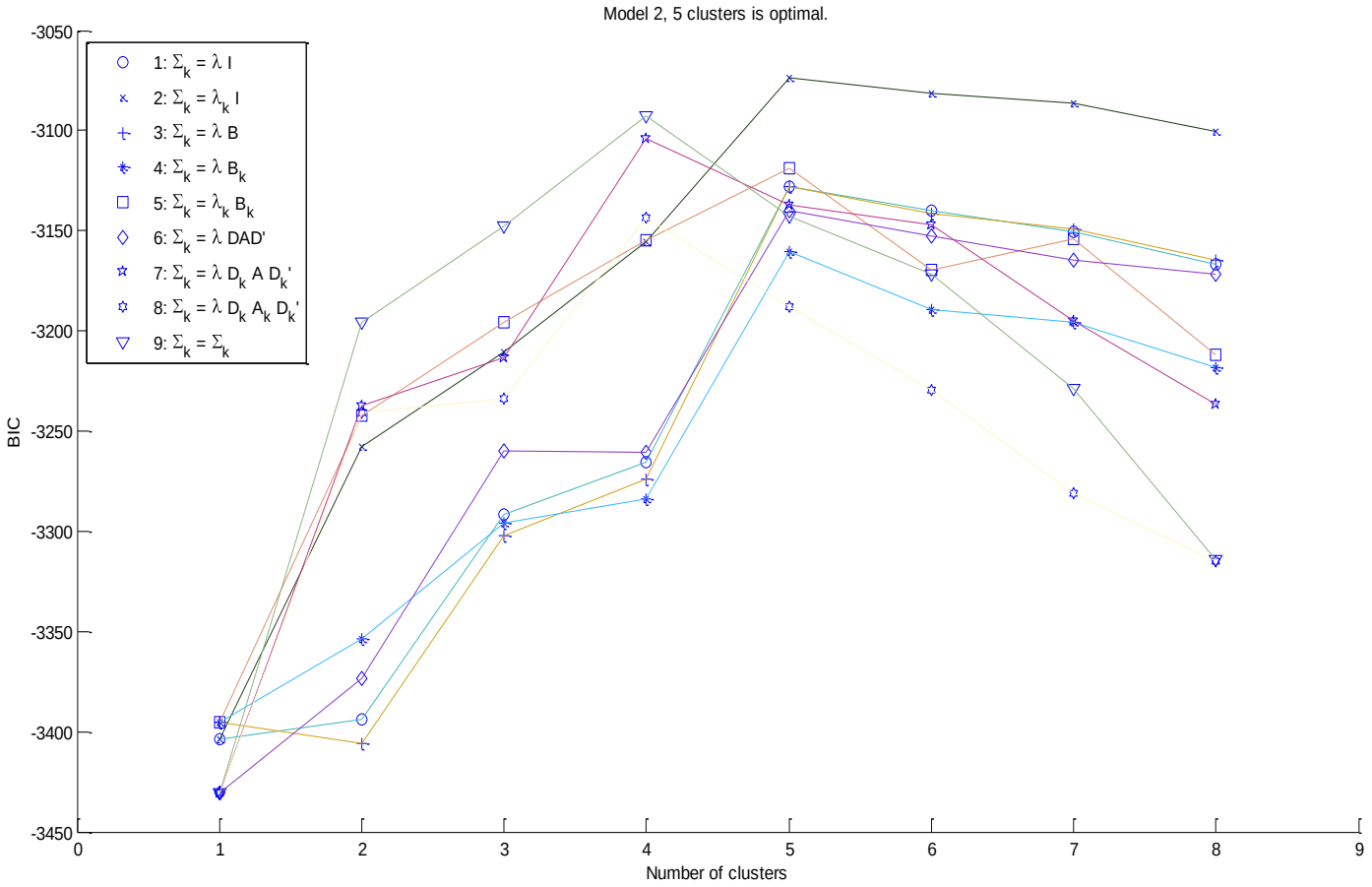
Step 4: Cluster Documents

- Model-Based Clustering
  - Based on Finite Mixtures and Expectation-Maximization (EM) algorithm
  - Provides a sequence of starting points for the EM algorithm applied to finite mixtures
  - Loop over different starting points based on:
    - Number of clusters
    - Models
  - Choose the 'best' model and number of clusters using the Bayesian Information Criterion – BIC

BLS

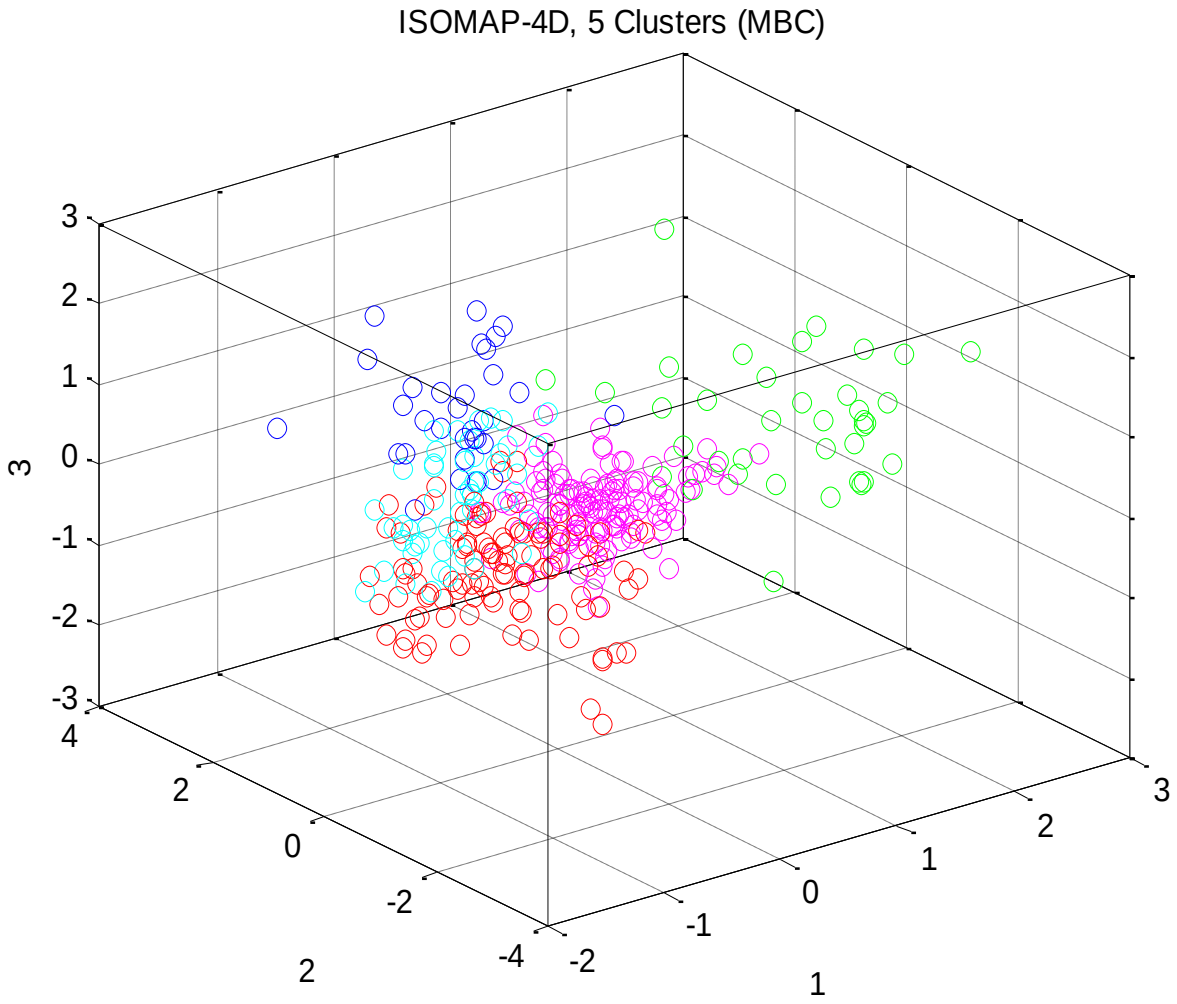
http://www.stat.washington.edu/raftery/Research/mbc.html

Bayesian Information Criterion – MBC



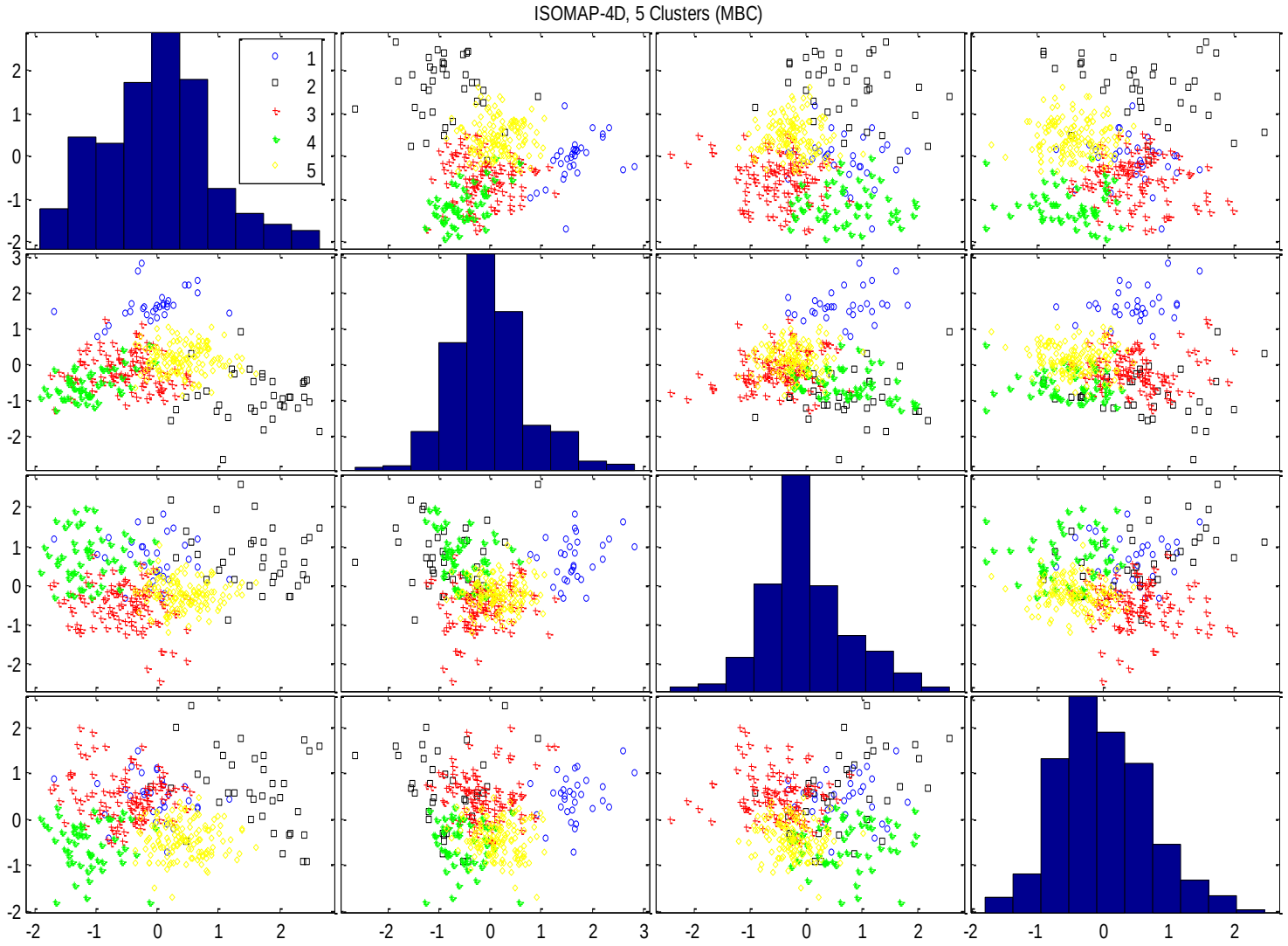
BLS

Model-Based Clustering



BLS

Model-Based Clustering



BLS

Step 5: Assess Clusters

Top 10 Words – MBC

Cluster 1 31 Docs	Cluster 2 37 Docs	Cluster 3 102 Docs	Cluster 4 51 Docs	Cluster 5 137 Docs
employee	employee	employee	employee	employee
truck	**machine**	december	**ladder**	november
trailer	**finger**	approximately	approximately	december
approximately	**saw**	**feet**	**roof**	**forklift**
november	**left**	**working**	**fell**	approximately
driver	**hand**	**fell**	working	**hospital**
cable	approximately	november	employees	**back**
pole	**blade**	employer	november	right
december	december	**tree**	december	left
struck	**cutting**	accident	ft	working

BLS

Refining the Analysis

- Compare clusters obtained using other approaches
  - K-means
  - Agglomerative
- Look for sub-clusters
- Adjust stop word list – redo analysis
- Try stemming and term weights
- Use alternative encoding – BPMs

BLS

Other Applications

- Cluster abstracts each month and look for trends in accidents
- Use clusters to help with coding and classification
- Use abstracts along with labels (e.g., event type) to create a classifier

BLS

Contact Information

Wendy Martinez

Director, Mathematical Statistics Research Center
Office of Survey Methods Research

www.bls.gov/osmr/home.htm
202-691-7400
martinez.wendy@bls.gov

BLS