Wilcoxon Signed-Rank Test to Compare Document Embedding Algorithms

Anton Chen

STAT 461 Statistical Inference II
The University of British Columbia

April 5, 2024

Motivation

- Sneak peak at how course material is applied
- Practice problem setup to apply methodology appropriately
- Comforting to know what we're learning has application :)

Context

High-dimensional data:

$$\mathbf{X} \in \mathbb{R}^{n \times d} \text{ where } d \text{ large}$$

- Increasingly common/relevant in modern day
- Broad applications, e.g. finance, genomics, ML

Context cont'd

Can be unwieldy:

- **Computational cost**
- Redundant and irrelevant features
- Curse of dimensionality (e.g. growing sparsity)

Context cont'd

Solution: **Dimensionality Reduction (DR)**

Definition (DR algorithm)

$$g = \left\{ g_{d'} : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d'} \text{ where } 0 < d' \leq d \right\}$$

- Transform $\mathbb{R}^d$ samples into $\mathbb{R}^{d'}$ where we specify output dimension d'
- Typically want
  1. $d' \ll d$
  2. "preserve" data quality/meaning, whatever that means

Context cont'd

E.g. the famous **PCA**

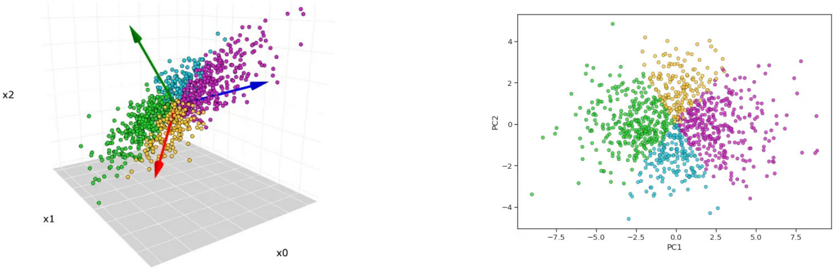

Figure: PCA $\mathbb{R}^3 \rightarrow \mathbb{R}^2$ [2]

Context cont'd

And many others, e.g.

- LDA,
- Isomap,
- t-SNE, and so on..

Point is, there are **many DR techniques**

Context cont'd

Similarly, there's many optimality criterion (**quality measures**):

Year	Name of the measure
1962	Sheppard Diagram (SD)
1964	Kruskal Stress Measure (S)
1969	Sammon Stress (S_S)
1988	Spearman's Rho (S_R)
1992	Topological Product (T_{Pr})
1997	Topological Function (T_F)
2000	Residual Variance (R_V)
2000	König's Measure (K_M)
2001	Trustworthiness & Continuity (T&C)
2003	Classification error rate
2006	Local Continuity Meta-Criterion (Q_k)
2006	Agreement Rate (A_R)/Corrected Agreement Rate (CA_R)
2007	Mean Relative Rank Errors (MRRE)
2009	Procrustes Measure (P_M)/Modified Procrustes Measure (P_{MC})
2009	Co-ranking Matrix (Q)
2011	Global Measure (Q_Y)
2011	The Relative Error (R_E)
2012	Normalization independent embedding quality assessment (NIEQA)

Figure: Well-known measures to evaluate DR algorithm quality, listed chronologically [3]

Context cont'd

For our purposes,

Definition (Quality Measure)

$$q \colon \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d'} \to [0, 1]$$

Given original dataset and reduced dataset, output quality score.

- Higher quality score is better
- Different DR algorithms optimize for different quality measures
- Can be combined: Gracia et al. take mean score of various well-regarded measures

Problem statement

Given dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$, DR algorithms g^1, g^2, and quality measure q, determine whether g^2 yields higher quality reductions than g^1.

Review

Wilcoxon Signed-Rank Test [1]

- Paired data (x_i, y_i) s.t. $x_i \neq y_i$ for $i = 1, \cdots, n'$
- Marginal distributions F, G s.t. $X_i \sim F$ and $Y_i \sim G$, need not be normal, regular, nor parametric
- Hypotheses

$$H_0 : F \equiv G \text{ v.s. } H_1 : F < G$$

Review cont'd

For $\Delta_i = |y_i - x_i|$, $\delta_i = \begin{cases} 1 & y_i > x_i \\ -1 & \text{o.w.} \end{cases}$, and

$$R_i = \sum_{j=1}^{n} \mathbb{1}\{\Delta_j < \Delta_i\} + \frac{1}{2}\sum_{j=1}^{n} \mathbb{1}\{\Delta_j = \Delta_i\} + \frac{1}{2},$$

the Wilcoxon signed-rank test statistic is

$$W_n = \sum_{i=1}^{n} \delta_i R_i$$

Review cont'd

with test and p-value

$$\phi(W_n) = \mathbb{1}\left\{W_n > z_{1-\alpha}\sqrt{\frac{1}{6}n(n+1)(2n+1)}\right\}, \qquad \text{(For size-}\alpha\text{ test)}$$
$$p = \mathbb{P}\left[W_n > w_0\right]. \qquad \text{(For observed } w_0\text{)}$$

- Recall asymptotic normality of W_n

Methodology (per Gracia et al.)

Step 1: choose range of target dimensions D:

- Commonly $D = \{2, 3, \cdots, d\}$

Methodology cont'd

Step 2: extract quality scores into $\mathbf{Q} \in [0,1]^{n \times 2}$ where

$$\mathbf{Q}_{ij} = q(\mathbf{X}, g_{d_i}^{j}(\mathbf{X})) \qquad \text{(For } d_i \in D\text{)}$$

Interpretation:

- jth column has the quality scores of DR algo j
- ith row are the paired quality scores of the DR algos compressing to d_i dimensions

Methodology cont'd

Step 3: apply Wilcoxon signed-rank on $\mathbf{Q}$

- Paired dataset $\mathbf{Q} = \begin{bmatrix} Q_{\cdot 1} & Q_{\cdot 2} \end{bmatrix}$
- F, G are quality score distributions of algos g^1, g^2 respectively (on $\mathbf{X}$)
- This tells us whether to reject H_0!

Example: Document Embedding for NLP

How do we represent a collection of documents as a matrix?

Example cont'd

E.g. Bag of Words [5]

Table: Document-term matrix $\mathbf{X} \in (\mathbb{Z}_+ \cup \{0\})^{n \times d}$

Doc/Term	**the**	**quick**	**brown**	**fox**	$\cdots$
Doc 1	1	1	1	1	$\cdots$
Doc 2	0	1	0	6	$\cdots$
Doc 3	0	100	1	6	$\cdots$
Doc 4	0	0	0	1	$\cdots$
Doc 5	0	0	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

- n documents, d dimensions (vocabulary size)
- d can be huge; think of the number of words in the English language!

Example cont'd

Some DR algorithms in the literature:

1. Doc2Vec: learn document representation via skip-grams
2. Latent Semantic Analysis (LSA): SVD on term-document matrix
3. More, but we'll focus on these 2

Example cont'd

Quality measure: based on downstream task

- Desirable characteristics: non-conflation, robustness against lexical ambiguity, demonstration of multifacetedness, reliability, etc. [4]

Example cont'd

Question: Doc2Vec claims to capture semantic information of the document. Does it represent documents better than LSA for classification?

Example cont'd

Experiment design:

- Dataset: `20newsgroups`, $\sim 20\,000$ documents, each with 1 of 20 topics (labels)
- DR algorithms: Doc2Vec, LSA
- Output dimensions: $50, 60, \cdots, 500$
- Quality measure: downstream classification training error via logistic regression
  - For illustrative purposes; not meant to be anything groundbreaking

Example cont'd

1. Packages

```
import numpy as np
from scipy.stats import wilcoxon
from sklearn.datasets import fetch_20newsgroups
from sklearn.decomposition import TruncatedSVD
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from gensim.models.doc2vec import Doc2Vec, TaggedDocument
```

Example cont'd

2. DR Algorithms (Doc2Vec)

```
def reduce_doc2vec(X, output_dim=50):
  tagged_data = [
      TaggedDocument(words=d.split(), tags=[str(i)]) for i, d in enumerate(X)
  ]

  model = Doc2Vec(
      tagged_data,
      vector_size=output_dim,
      window=5,
      min_count=5,
      epochs=3,
  )

  X_doc2vec = np.array([model.infer_vector(doc.split()) for doc in X])

  return X_doc2vec
```

Example cont'd

2. DR Algorithms (LSA)

```
def reduce_lsa(X, output_dim=50):
  # Convert text documents to a document-term matrix
  vectorizer = CountVectorizer()
  X_counts = vectorizer.fit_transform(X)

  # Apply SVD
  lsa = TruncatedSVD(n_components=output_dim)
  X_lsa = lsa.fit_transform(X_counts)

  return X_lsa
```

Example cont'd

3. Get quality scores

```
def get_quality_scores(X, y, output_dims):
  quality_scores = np.zeros((len(output_dims), 2))

  for i, output_dim in enumerate(output_dims):
    X_lsa = reduce_lsa(X, output_dim)
    X_doc2vec = reduce_doc2vec(X, output_dim)

    quality_scores[i][0] = fit_logistic(X_lsa, y)
    quality_scores[i][1] = fit_logistic(X_doc2vec, y)

  return quality_scores
```

```
# Get quality scores
output_dims = [i for i in range(50, 501, 10)]
quality_scores = get_quality_scores(X_train, y_train, output_dims)
```

Example cont'd

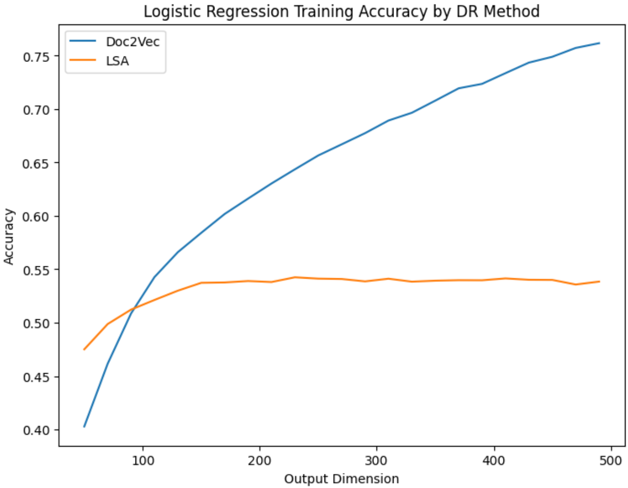

Figure: Quality score samples

Example cont'd

4. Wilcoxon Signed-Rank Test

```
# Hypothesis test
statistic, p_value = wilcoxon(
    quality_scores[:, 0],
    quality_scores[:, 1],
)

print("Wilcoxon Signed-Rank Test:")
print("Test Statistic:", statistic)
print("p-value:", p_value)

alpha = 0.05
if p_value < alpha:
  print("Reject H_0")
else:
  print("Do not reject H_0")
```

tl;dr

- Dimensionality Reduction (DR), many techniques exist, evaluate with quality measure
- Wilcoxon signed rank test: compare quality scores of two DR algorithms
- Example: document embeddings for NLP

References

[1] Jiahua Chen. *STAT 460/560 + 461/561: Statistical Inference I & II*. 2023/2024.

[2] Casey Cheng. *Principal Component Analysis (PCA) Explained Visually with Zero Math*. Published in Towards Data Science, Feb 3. Towards Data Science. 2022. URL: https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d.

[3] Antonio Gracia et al. "A methodology to compare dimensionality reduction algorithms in terms of loss of quality". In: *Information Sciences* 270 (2014), pp. 1–27.

[4] Bin Wang et al. "Evaluating word embedding models: Methods and experimental results". In: *APSIPA transactions on signal and information processing* 8 (2019), e19.

[5] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. "Understanding bag-of-words model: a statistical framework". In: *International journal of machine learning and cybernetics* 1 (2010), pp. 43–52.