Latent Semantic Analysis for Finding Relevant Biomedical literature

Santina Lin,[1,2] Jake Lever,[1,2] Sita Gakkhar,[2] Steven Jones[2,3]

1. Bioinformatics Training Program, UBC | 2. Michael Smith Genome Sciences Centre | 3. Department of Medical Genetics, UBC || Vancouver, British Columbia, Canada

Overview

There are ~25 million records in the PubMed database. This astounding number makes it difficult for scientists to find publications relevant to their work, and the pace of biomedical research is increasing. Hence, finding better ways to retrieve relevant scientific papers is important. In this study, we look at how to use latent semantic analysis (LSA)—an application of singular value decomposition (SVD)—on finding relationships in a set of documents and terms to retrieve similar biomedical papers.

Conclusion

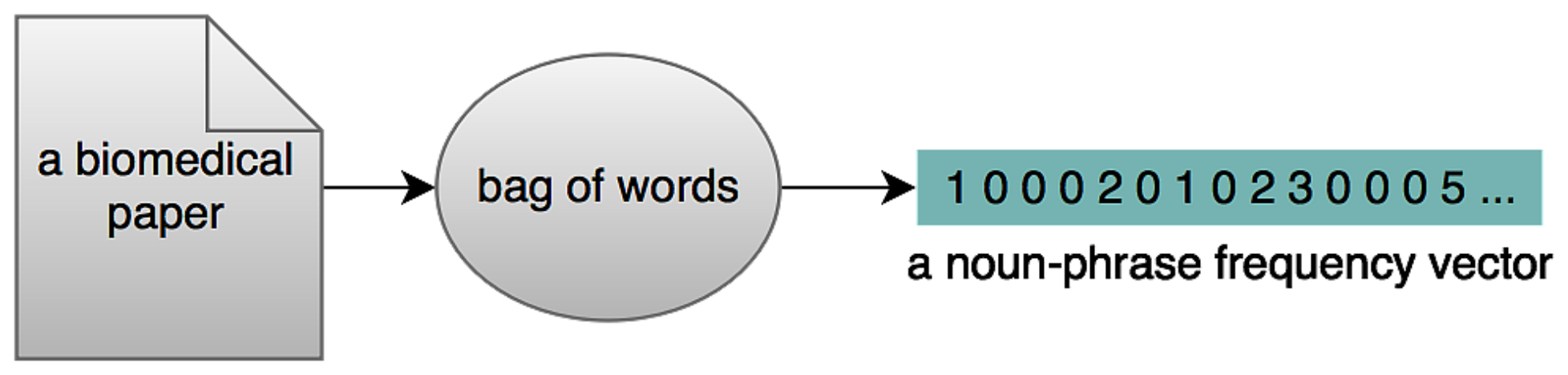
We demonstrated that SVD, a machine learning method, can generalize concepts in biomedical literature and thus find similar papers. After using different parameters of SVD to retrieve related biomedical papers that are suggested by PubMed, we found that **using a term frequency–inverse document frequency (TF–IDF) matrix in combination with cosine distance yields the best precision**. The next step is to investigate how SVD would perform on a much larger corpus (*i.e.* millions of papers).

Background

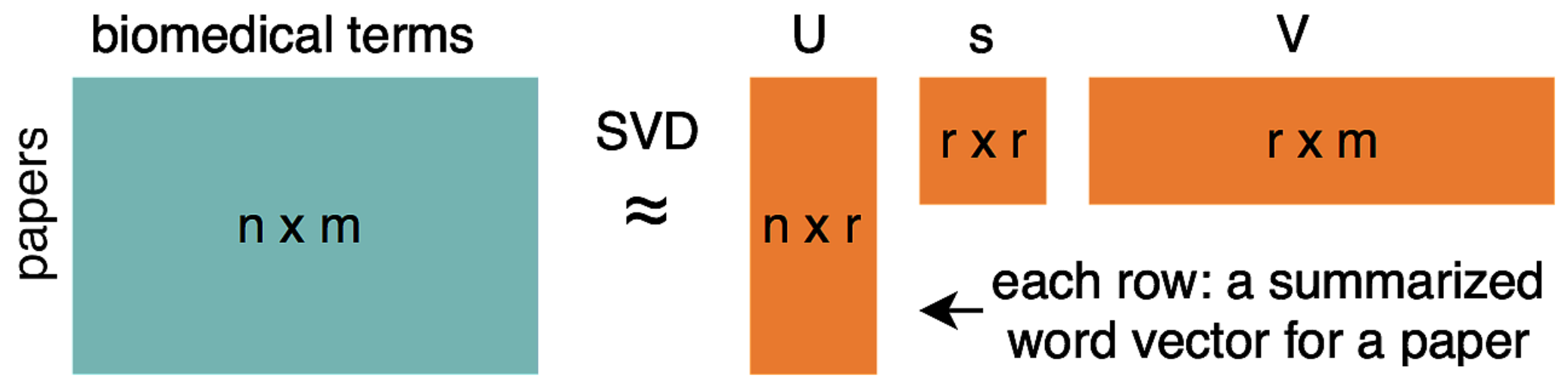
Many frameworks that retrieve related articles require information that is not always available, such as full text, MeSH headings, and citations.

For example, PubMed identifies a number of related articles using the manually assigned MeSH headings. As a result, their method is not scalable.

SVD is a matrix factorization method that decomposes a matrix into a set of uncorrelated components. As used in PCA, it identifies dimensions on which the data exhibits the most variance and represents the simplified and salient structures underlying the data.



A paper-term matrix is constructed by matching papers' words to a list of biomedical terms. The occurrences are recorded in rows of the matrix.



SVD decomposes the matrix into three matrices. Rows in U are summarized word vectors, which are used to measure distances between papers.

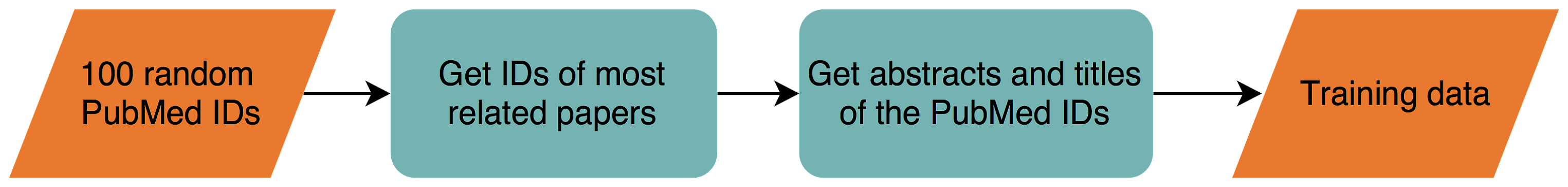
SVD has been widely applied in recommendation systems such recommenders for movies and products.

We want to use SVD to recommend similar articles because it can generalize the patterns in a large corpus of text and thus address some problems, such as term independence, synonymy, and noise, in other retrieval methods. For example, two papers using different synonyms describing the same topic would be seemed as dissimilar by a naive method, but SVD would identify them as similar by recognizing that those vocabularies often occur together in other papers.

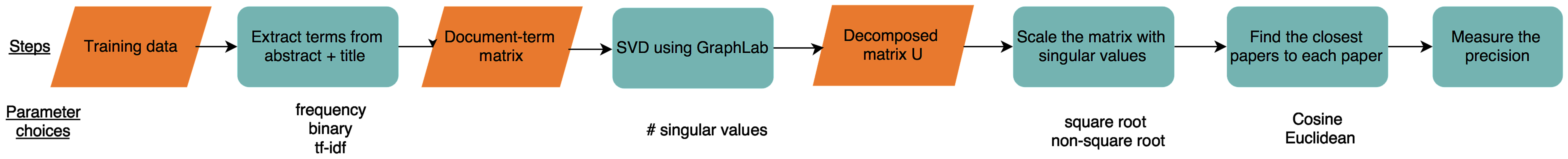
There are parameters to optimize when performing SVD, such as the matrix type and the distance function. This study explore different combination of parameters for finding related biomedical literature.

Methods

1. Generating training data using Entrez.Bio:
   A python library for querying NCBI databases.


100 random PubMed IDs
Get IDs of most related papers
Get abstracts and titles of the PubMed IDs
Training data

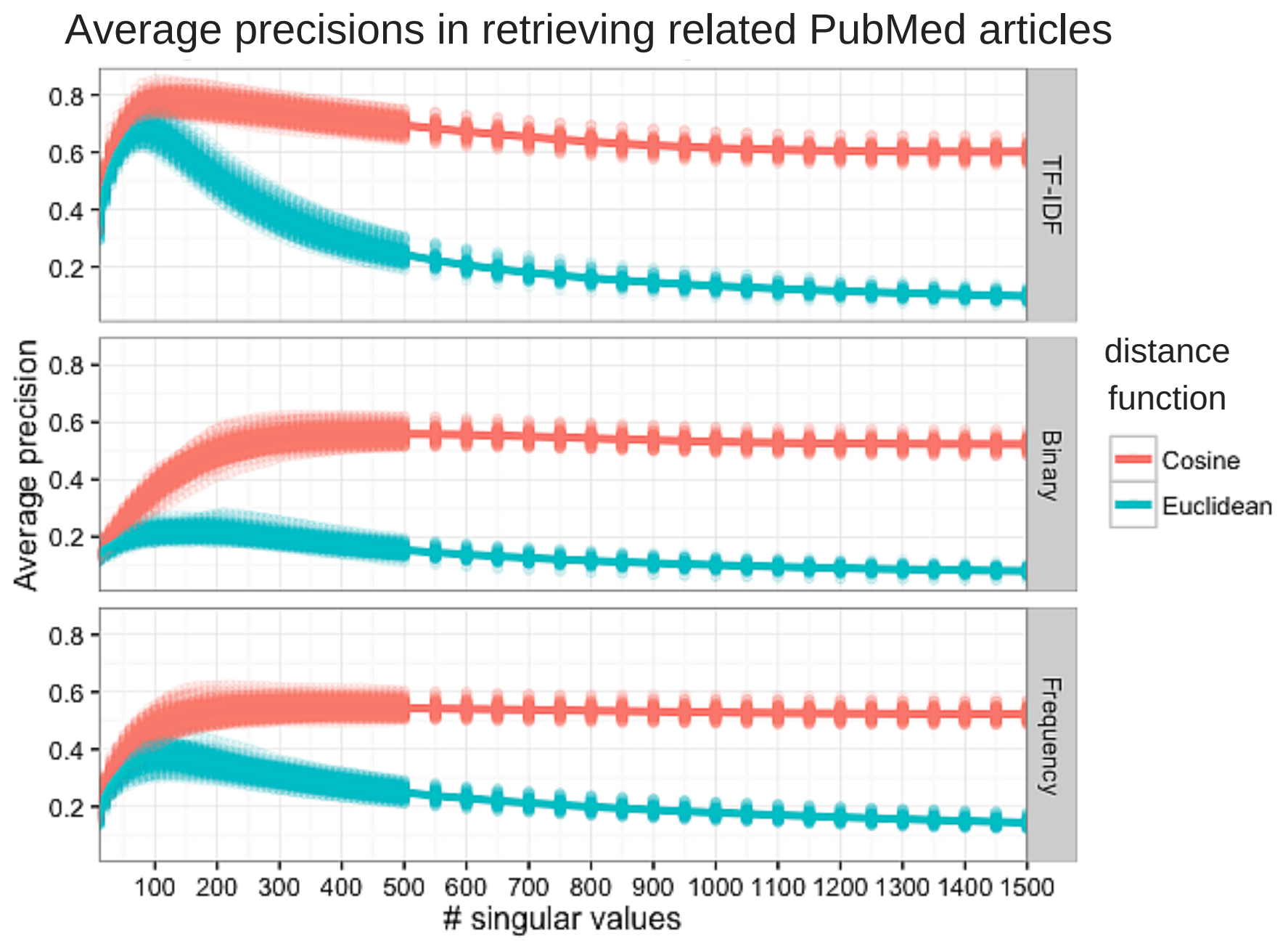
2. Finding the best parameters for SVD by evaluating the precisions of different combinations of parameters.



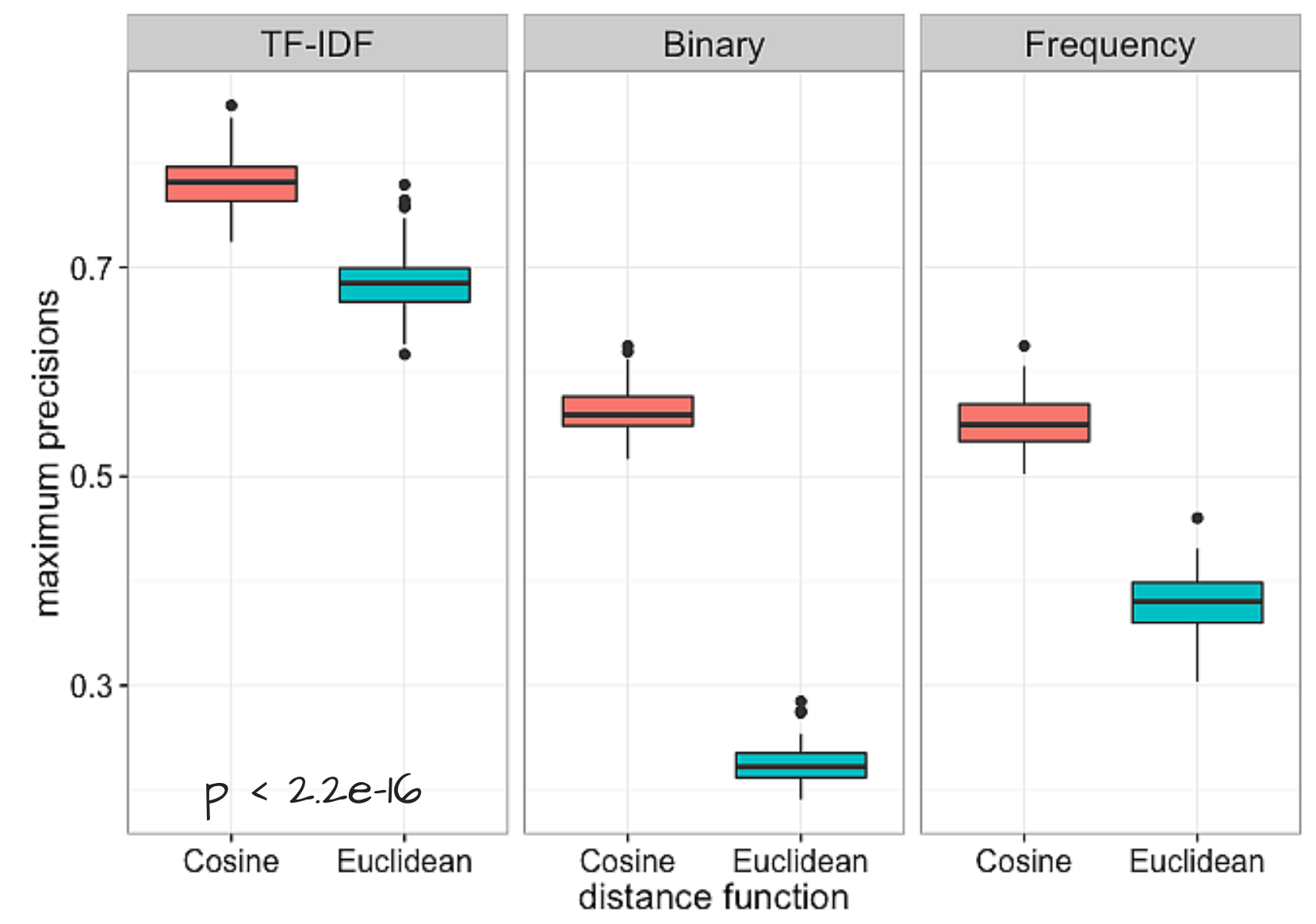
Results

1. TF–IDF and cosine distance function work best

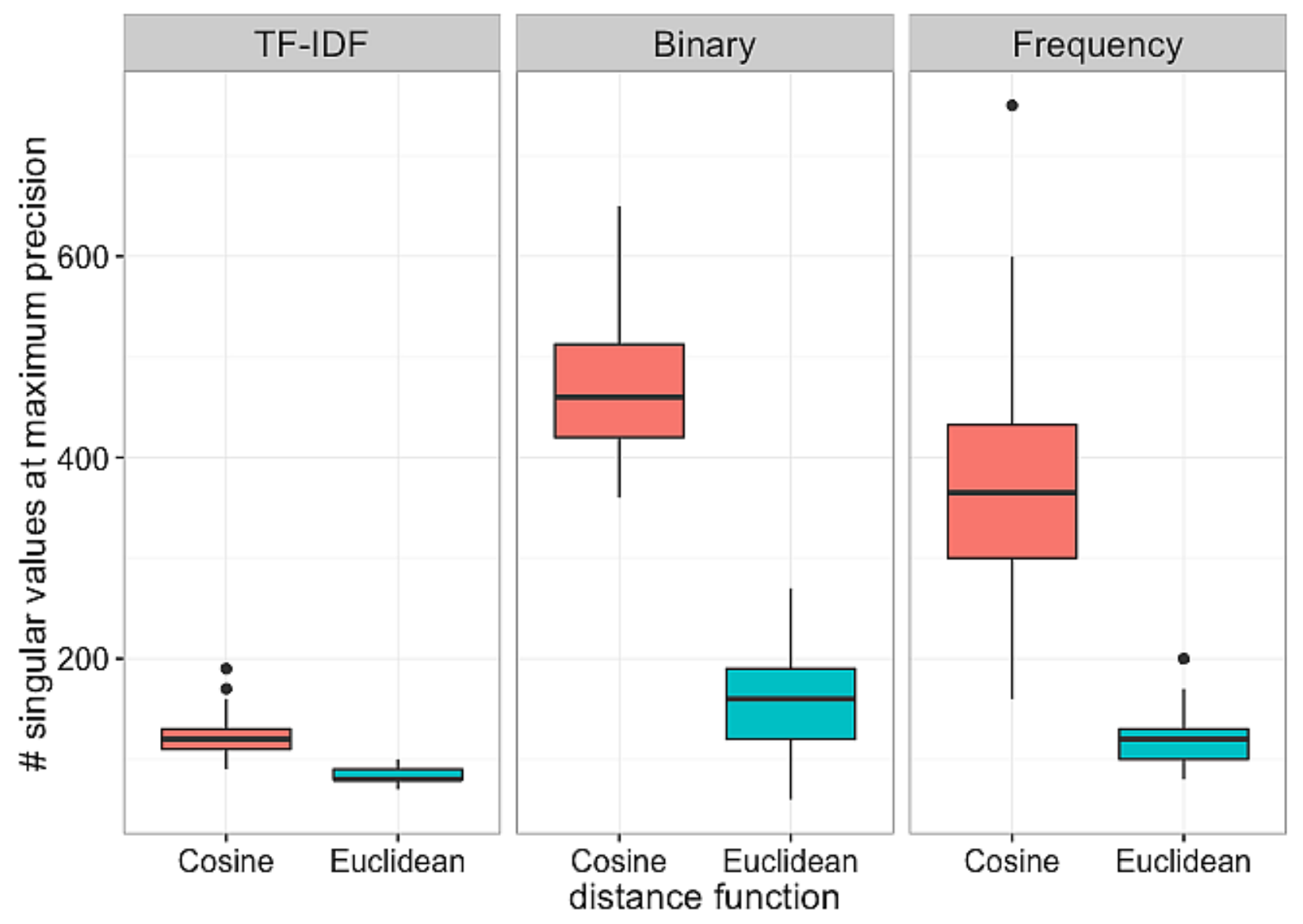
Average precisions in retrieving related PubMed articles



Maximum precisions achieved using different matrix type and distance function



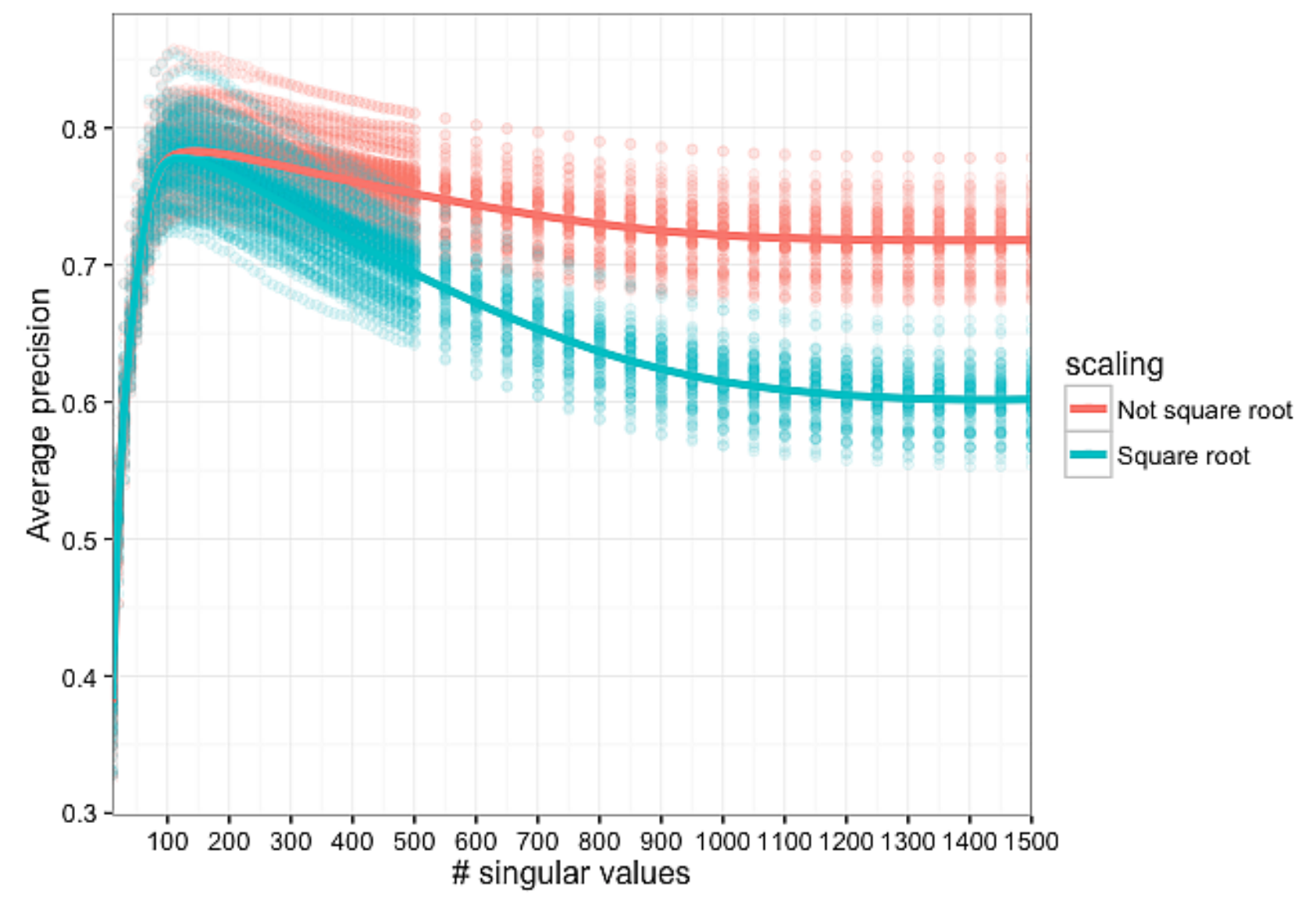
Numbers of singular values at maximum precisions achieved by different matrix type and distance function



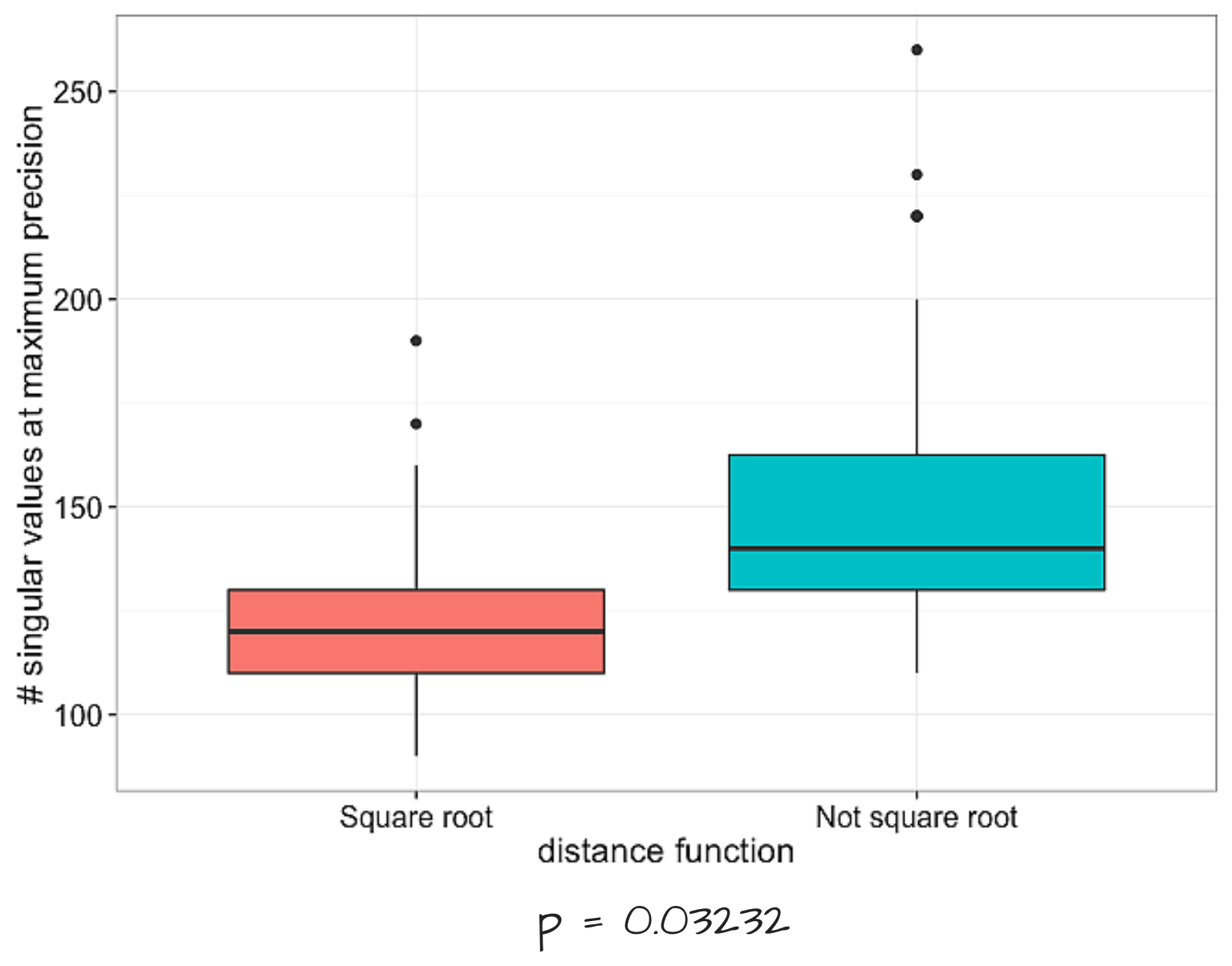
The curves peak and then taper. This indicates that after a number of singular values, SVD has generalized enough concepts to make the best prediction before the inclusion of noise starts to affect the precision.

2. Scale the matrix with singular values might result in higher precisions

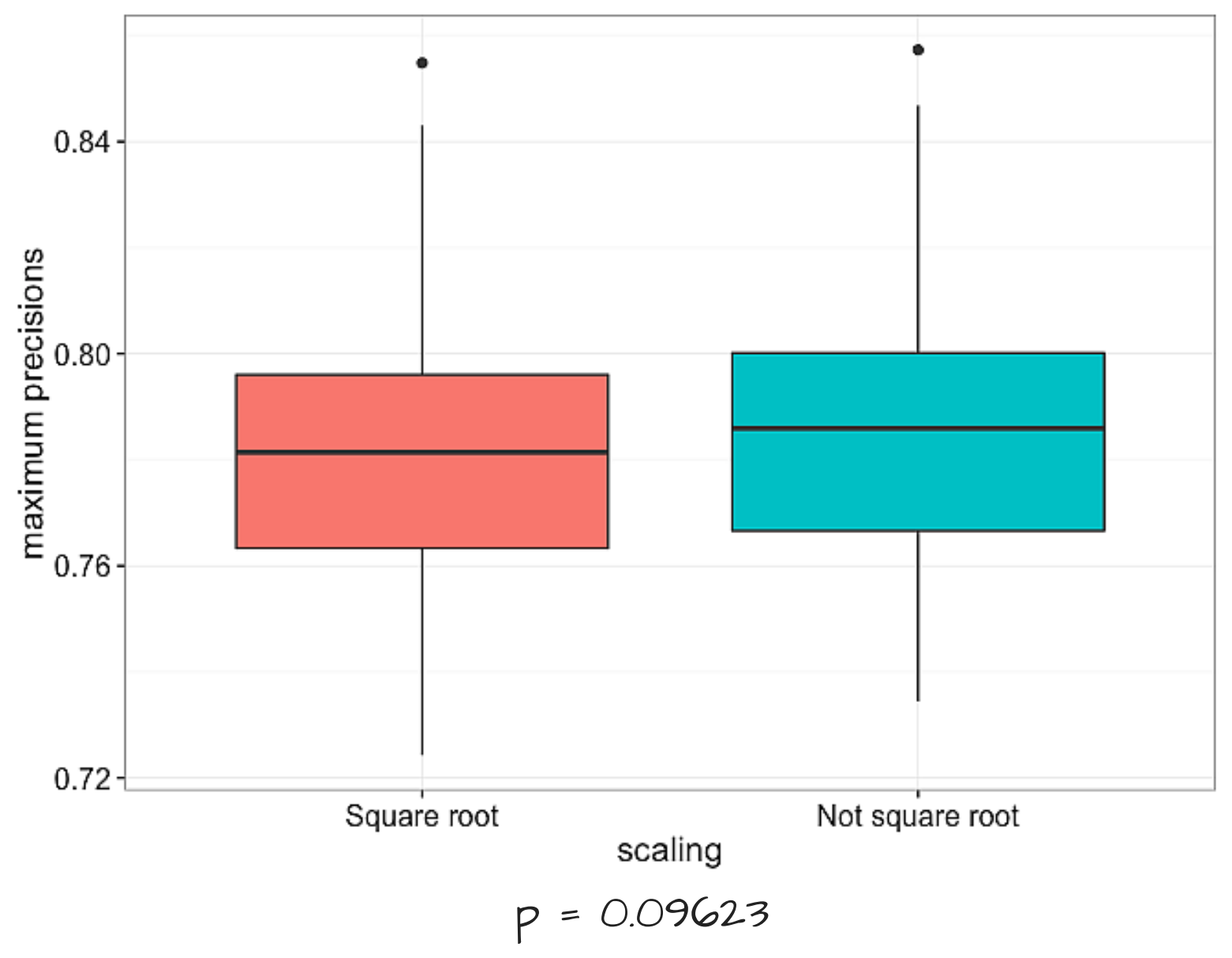
Average precisions in retrieving related PubMed articles



Maximum precisions achieved using different scaling



Numbers of singular values at maximum precisions achieved by different scaling method



Acknowledgments

I would like to thank the Jones lab members for the valuable discussions; Jake and Sita for their help setting up the software environment and creating a list of biomedical terms; Inanc Birol, Sara Mostafavi, and Steven Jones for their committee memberships; and the CIHR Strategic Training Scholarship.

Related Works

Chen H, et al. Front. Physiol., 2013 January 30
Entrez Programming Utilities Help [Internet]. Bethesda (MD): NCBI (US); 2010-.
Ekstrand MD, et al. Foundations and Trends in Human–Computer Interaction. 2010