Text Mining for the Social Sciences
Lecture 4: Latent Semantic Analysis

Stephen Hansen

Introduction

Last time we considered analysis of the document-term matrix $\mathbf{X}$ for information retrieval.

The vector space model treats each term in the vocabulary as an independent source of variation.

Properties of language make this assumption quite restrictive: synonomy and polysemy.

The introduction of latent variables provides a way of accounting for correlations among words.

Rest of course focuses on this.

Synonomy

The same underlying concept can generally be described by many different words.

Words associated with the theme of education are 'school', 'university', 'college', 'teacher', 'professor', etc.

Consider the following two documents

school	university	college	teacher	professor
0	5	5	0	2
school	university	college	teacher	professor
10	0	0	4	0

How big is their cosine similarity?

Polysemy

Polysemy refers to the same word having multiple meanings in different contexts.

Consider the following two documents

tank	seal	frog	animal	navy	war
5	5	3	2	0	0

tank	seal	frog	animal	navy	war
5	5	0	0	4	3

How related are these documents? How large is their cosine similarity?

Latent Variables

Recall the example from last lecture of the dictionary $\mathfrak{D} = \{\text{labor}, \text{wage}, \text{employ}\}$.

The implicit assumption is that these words map back into an underlying topic $\{\text{labor}, \text{wage}, \text{employ}\} \rightarrow \{\text{labor markets}\}$.

We cannot observe the topics in text, only observe the words that those topics tend to generate.

A natural way forward is to model topics with latent variables.

Features of Latent Variable Models

Latent variable models generally share the following features:

1. Associate each word in the vocabulary to any given latent variable (synonomy).

2. Allow each word to have associations with multiple topics (polysemy).

3. Associate each document with topics.

Give documents a representation in a latent, more accurate, semantic space rather than the raw vocabulary space.

Allow algorithm to find best association between words and latent variables $\Rightarrow$ no pre-defined word lists or labels. Can be more objective than dictionary methods.

Latent Semantic Analysis

The first latent variable model we'll consider is latent semantic analysis (a.k.a. latent semantic indexing).

Original reference is Deerwester et. al. (1990) on reading list.

In common with this literature, for the purposes of this lecture treat $\mathbf{X}$ as a term-document matrix rather than a document-term matrix.

Influential approach in information retrieval, precursor of later, more probabilistic approaches.

General Idea

Recall principal components analysis: project covariance matrix of observed variables onto new dimensions such that dimensions are orthogonal and ranked by the variance they explain.

One can then represent the data in terms of the first k principal components.

This represents the meaningful variation in the data in a lower dimensional space, and strips out random noise.

Text Mining Context

Consider the variation of terms across documents.

Some of this variation will come from the fact that different documents have different content, and some from idiosyncratic vocabulary choices.

We'd like to do something like principal components analysis:

1. Find "content" dimensions that explains covariance of terms across documents separately from "noise" dimensions that don't.
2. Then approximate $\mathbf{X}$ with a lower-ranked reconstruction $\hat{\mathbf{X}}$ using the content dimensions.

Even though $\hat{\mathbf{X}}$ is an approximation to $\mathbf{X}$, the hope is that it provides a *more accurate* representation of each document's content.

LSA uses a generalization of the eigenvalue decomposition for square matrices called *singular value decomposition*.

Singular Value Decomposition

Assume $D > V$ and that $\text{rank}(\mathbf{X}) = V$.

Proposition

The term-document matrix can be written $\mathbf{X} = \mathbf{A}\mathbf{\Sigma}\mathbf{B}^T$ *where* $\mathbf{A}$ *is a* $V \times V$ *orthogonal matrix,* $\mathbf{B}$ *is a* $D \times D$ *orthogonal matrix, and* $\mathbf{\Sigma}$ *is a* $V \times D$ *matrix where* $\mathbf{\Sigma}_{ii} = \sigma_i$ *with* $\sigma_i \geq \sigma_{i+1}$ *for* $1 \geq i \geq V - 1$ *and* $\mathbf{\Sigma}_{ij} = 0$ *for all* $i \neq j$.

Singular Value Decomposition

Assume $D > V$ and that $\text{rank}(\mathbf{X}) = V$.

Proposition

The term-document matrix can be written $\mathbf{X} = \mathbf{A}\mathbf{\Sigma}\mathbf{B}^T$ *where* $\mathbf{A}$ *is a* $V \times V$ *orthogonal matrix,* $\mathbf{B}$ *is a* $D \times D$ *orthogonal matrix, and* $\mathbf{\Sigma}$ *is a* $V \times D$ *matrix where* $\mathbf{\Sigma}_{ii} = \sigma_i$ *with* $\sigma_i \geq \sigma_{i+1}$ *for* $1 \geq i \geq V - 1$ *and* $\mathbf{\Sigma}_{ij} = 0$ *for all* $i \neq j$.

Some terminology:

- Columns of $\mathbf{A}$ are called left singular vectors.
- Columns of $\mathbf{B}$ are called right singular vectors.
- The diagonal terms of $\mathbf{\Sigma}$ are called singular values.

Shape of $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \begin{bmatrix} \overbrace{\begin{matrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_V \end{matrix}}^{V} & \overbrace{\begin{matrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{matrix}}^{D-V} \end{bmatrix}$$

A more parsimonious decomposition drops the last $D - V$ columns of $\mathbf{\Sigma}$ and the last $D - V$ rows of $\mathbf{B}^T$. Goes by a variety of names (thin, reduced, etc.).

Relationship to Eigenvalue Decomposition

By the above result, we have that

$$\mathbf{X}\mathbf{X}^T = \mathbf{A}\mathbf{\Sigma}\mathbf{B}^T\mathbf{B}\mathbf{\Sigma}^T\mathbf{A}^T = \mathbf{A}\left(\mathbf{\Sigma}\mathbf{\Sigma}^T\right)\mathbf{A}^T \Rightarrow \left(\mathbf{X}\mathbf{X}^T\right)\mathbf{A} = \mathbf{A}\left(\mathbf{\Sigma}\mathbf{\Sigma}^T\right)$$

where

$$\mathbf{\Sigma}\mathbf{\Sigma}^T = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_V^2 \end{bmatrix}.$$

This implies that

$$\left(\mathbf{X}\mathbf{X}^T\right)\mathbf{a}_i = \sigma_i^2\mathbf{a}_i$$

for every column $\mathbf{a}_i$ of $\mathbf{A}$ and every singular value σ_i.

Result

For $i = 1, \dots, V$, $\mathbf{a}_i$ is an eigenvector of $\mathbf{X}\mathbf{X}^T$ with associated eigenvalue σ_i^2.

Relationship to Eigenvalue Decomposition

Using the exact same logic, one can show that

Result

For $i = 1, \ldots, V$, $\mathbf{b}_i$ is an eigenvector of $\mathbf{X}^T\mathbf{X}$ with associated eigenvalue σ_i^2.

Note that for $i = V + 1, \ldots, D$, $\mathbf{b}_i$ is an eigenvector of $\mathbf{X}^T\mathbf{X}$ with a zero eigenvalue. $\mathbf{X}^T\mathbf{X}$ does not have full rank.

Relationship to Principal Components

Suppose we treat columns as observations and rows as covariates, and imagine that the rows of $\mathbf{X}$ have been demeaned.

Then the covariance matrix is a linear scaling of $\mathbf{X}\mathbf{X}^T$.

The columns of $\mathbf{A}$ are then principal components.

Similarly, the columns of $\mathbf{B}$ are principal components treating rows as observations and columns as covariates.

SVD is thus equivalent to principal components if data is centered.

Role of Term-Document Matrix

The term-document $\mathbf{X}$ to which we apply SVD as part of LSA can take several forms (incidence matrix, raw frequency counts, tf-idf weighted counts, etc.).

Suppose $\mathbf{X}$ is an incidence matrix:

1. The (i, j)th entry of $\mathbf{X}\mathbf{X}^T$ gives the number of documents in which terms i and j both appear.
2. The (i, j)th entry of $\mathbf{X}^T\mathbf{X}$ gives the number of terms that documents i and j share.

More generally, $\mathbf{X}\mathbf{X}^T$ measures term overlap and $\mathbf{X}^T\mathbf{X}$ measures document overlap.

Left singular vectors of $\mathbf{X}$ associate terms to topics, and right singular vectors associate documents to topics.

Approximating Term-Document Matrix

Recall the goal is to approximate $\mathbf{X}$ with a matrix $\hat{\mathbf{X}}$ that represents terms and documents in a latent space of thematic topics while stripping out noise.

In the spirit of PCA, relate singular vectors of $\mathbf{X}$ to magnitude of singular values.

Let $\mathbf{\Sigma}_k$ be the diagonal matrix formed by replacing $\mathbf{\Sigma}_{ii} = 0$ for $i = k+1, \ldots, V$.

Let $\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}}$. We can quantify the approximation error using the Frobenius norm

$$\|E\|_F = \sqrt{\sum_{v=1}^{V} \sum_{d=1}^{D} e_{vd}^2}.$$

Eckart-Young Theorem

Proposition

The solution to the problem of choosing $\hat{\mathbf{X}}$ *to minimize* $\|E\|_F$ *subject to the constraint that* $\text{rank}(\hat{\mathbf{X}}) = k$ *is given by* $\mathbf{X}_k^* = \mathbf{A}\mathbf{\Sigma}_k\mathbf{B}^T$.

Best low-rank approximation represents $\mathbf{X}$ in terms of the singular vectors that multiply largest singular values.

Note that $\mathbf{X}_k^*$ still has size $V \times D$, but its rank is lower.

The "effective" dimensionality of $\mathbf{A}$ is reduced from $V \times V$ to $V \times k$, and of $\mathbf{B}$ from $D \times V$ to $D \times k$.

Comparing Documents

Latent semantic indexing refers to comparing the similarity of documents using columns of $\mathbf{X}_k^*$.

Equivalently, we can treat document i as lying in a k-dimensional vector space with coordinates

$$\vec{d}_i = (\sigma_1 b_{i1}, \ldots, \sigma_k b_{ik}) .$$

We can then compare $\vec{d}_i$ and $\vec{d}_j$ as before, for example with cosine similarity.

If the k retained singular values capture thematic content, then we should get a more accurate measure of similarity than in the standard vector space.

Treating Queries

Recall the task of ranking documents relative to an out-of-sample document.

How can we transform a query into the same latent space as within-sample documents?

Suppose the approximation is perfect. Then

$$\mathbf{X} = \mathbf{A}\mathbf{\Sigma}_k\mathbf{B}^T \Rightarrow \mathbf{B}^T = \mathbf{\Sigma}_k^{-1}\mathbf{A}^T\mathbf{X}.$$

We can therefore transform any document $\vec{d}$ into the latent space via the transformation $\vec{d'} = \mathbf{\Sigma}_k^{-1}\mathbf{A}^T\vec{d}$, and then compute its cosine similarity with respect to the within-sample documents.

Comparing Terms

We might also be interested in how close any two terms are to assess synonomy.

Can compute similarity of rows of $\mathbf{X}_k^*$.

Equivalently, we can treat term i as lying in a k-dimensional vector space with coordinates

$$\vec{t_i} = (\sigma_1 a_{i1}, \ldots, \sigma_k a_{ik}) .$$

Example

Suppose the term-document matrix is given by

$$\mathbf{X} = \begin{array}{c} \\ \text{car} \\ \text{automobile} \\ \text{ship} \\ \text{boat} \end{array} \begin{array}{c} \begin{array}{cccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \end{array} \\ \begin{bmatrix} 10 & 5 & 0 & 0 & 1 & 0 \\ 0 & 5 & 14 & 2 & 0 & 0 \\ 1 & 1 & 0 & 10 & 20 & 2 \\ 0 & 1 & 0 & 5 & 21 & 7 \end{bmatrix} \end{array}$$

Cosine Similarity between Documents

	d_1	d_2	d_3	d_4	d_5	d_6
d_1	1	·	·	·	·	·
d_2	0.70	1	·	·	·	·
d_3	0.00	0.69	1	·	·	·
d_4	0.08	0.30	0.17	1	·	·
d_5	0.10	0.21	0.00	0.92	1	·
d_6	0.02	0.17	0.00	0.66	0.88	1

SVD

Can use numpy to compute SVD.

$$\mathbf{A} = \begin{bmatrix} 0.0503 & 0.2178 & -0.9728 & 0.0595 \\ 0.0380 & 0.9739 & 0.2218 & 0.0291 \\ 0.7024 & -0.0043 & -0.0081 & -0.7116 \\ 0.7088 & -0.0634 & 0.0653 & 0.6994 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 0.0381 & 0.1435 & -0.8931 & -0.02301 & 0.3765 & 0.1947 \\ 0.0586 & 0.3888 & -0.3392 & 0.0856 & -0.7868 & -0.3222 \\ 0.0168 & 0.9000 & 0.2848 & 0.0808 & 0.3173 & 0.0359 \\ 0.3367 & 0.1047 & 0.0631 & -0.7069 & -0.2542 & 0.5542 \\ 0.9169 & -0.0792 & 0.0215 & 0.1021 & 0.1688 & -0.3368 \\ 0.2014 & -0.0298 & 0.0404 & 0.6894 & -0.2126 & 0.6605 \end{bmatrix}$$

Singular Values

The singular values are $\boldsymbol{\sigma} = (31.61, 15.14, 10.90, 5.03)$.

If we take $k = 2$, we obtain

$$\mathbf{X}_k^* = \begin{array}{c} \\ \text{car} \\ \text{automobile} \\ \text{ship} \\ \text{boat} \end{array} \begin{array}{c} \begin{array}{cccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \end{array} \\ \begin{bmatrix} 0.5343 & 1.3765 & 2.9969 & 0.8817 & 1.1978 & 0.2219 \\ 2.1632 & 5.8077 & 13.2992 & 1.9509 & 0.0670 & 0.1988 \\ 0.8378 & 1.2765 & 0.3153 & 7.4715 & 20.3682 & 4.4748 \\ 0.7169 & 0.9399 & 0.4877 & 7.4456 & 20.6246 & 4.5423 \end{bmatrix} \end{array}$$

Cosine Similarity between Documents

	d_1	d_2	d_3	d_4	d_5	d_6
d_1	1	·	·	·	·	·
d_2	0.97	1	·	·	·	·
d_3	0.91	0.97	1	·	·	·
d_4	0.60	0.43	0.23	1	·	·
d_5	0.45	0.26	0.05	0.98	1	·
d_6	0.47	0.29	0.07	0.98	0.99	1

Application: Transparency

How transparent should a public organization be?

Benefit of transparency: accountability.

Costs of transparency:

1. Direct costs
2. Privacy
3. Security
4. Worse behavior $\rightarrow$ "chilling effect"

Transparency and Monetary Policy

Mario Draghi (2013): "It would be wise to have a richer communication about the rationale behind the decisions that the governing council takes."

Table: Disclosure Policies as of 2014

	Fed	BoE	ECB
Minutes?	✓	✓	X
Transcripts?	✓	X	X

Natural Experiment

FOMC meetings were recorded and transcribed from at least the mid-1970's in order to assist with the preparation of the minutes.

Committee members unaware that transcripts were stored prior to October 1993.

Greenspan then acknowledged the transcripts' existence to the Senate Banking Committee, and the Fed agreed:

1. To begin publishing them with a five-year lag.
2. To publish the back data.

"All the News That's Fit to Print"

The New York Times

VOL. CLXIII . . . No. 56,420 © 2014 The New York Times SATURDAY, FEBRUARY 22, 2014

Fed Misread Fiscal Crisis, Records Show

After Caution in 2008, Series of Bold Steps

By BINYAMIN APPELBAUM

WASHINGTON — On the morning after Lehman Brothers filed for bankruptcy in 2008, most Federal Reserve officials still believed that the American economy would keep growing despite the metastasizing financial crisis.

The Fed's policy-making committee voted unanimously against bolstering the economy by cutting interest rates, and several officials praised what they described as the decision to let Lehman fail, saying it would help to restore a sense of accountability on Wall Street.

James Bullard, president of the Federal Reserve Bank of St. Louis, urged his colleagues "to wait for some time to assess the impact of the Lehman bankruptcy filing, if any, on the national economy," according to transcripts of the Fed's 2008 meetings that it published on Friday.

DETROIT OUTLINES MAP TO SOLVENCY, STRESSING REPAIR

WAY OUT OF BANKRUPTCY

Balancing Act Worries Banks and Angers Retirees in City

By MONICA DAVEY and MARY WILLIAMS WALSH

DETROIT — Seven months after this city entered bankruptcy, its leaders on Friday presented a federal judge with the first official road map to Detroit's future — documents designed to show how it aims to settle its $18 billion debt to creditors and make itself livable again.

But the proposal is less a vision for a brand-new city than a repair estimate for the old one. It is a document designed by lawyers and bankruptcy experts to find ways to pay off more than 100,000 creditors and then budget money over a period of years to create a

Deal Signed in Ukraine, but Shows Stra

Greenspan's View on Transparency

"**A considerable amount of free discussion and probing questioning by the participants of each other and of key FOMC staff members takes place.** In the wide-ranging debate, new ideas are often tested, many of which are rejected ... **The prevailing views of many participants change as evidence and insights emerge.** This process has proven to be a very effective procedure for gaining a consensus ... It could not function effectively if participants had to be concerned that their half-thought-through, but nonetheless potentially valuable, notions would soon be made public. **I fear in such a situation the public record would be a sterile set of bland pronouncements scarcely capturing the necessary debates which are required of monetary policymaking.**"

Measuring Disagreement

Acosta (2014) uses LSA to measure disagreement before and after transparency.

For each member i in each meeting t, let $\vec{d}_{it}$ be member i's words.

Let $\vec{d}_{-i,t} = \sum_i \vec{d}_{it} - \vec{d}_{it}$ be all other members' words.

Quantity of interest is the similarity between $\vec{d}_{it}$ and $\vec{d}_{-i,t}$.

Total set of documents—$\vec{d}_{it}$ and $\vec{d}_{-i,t}$ for all meetings and speakers—is 6,152.

Singular Values

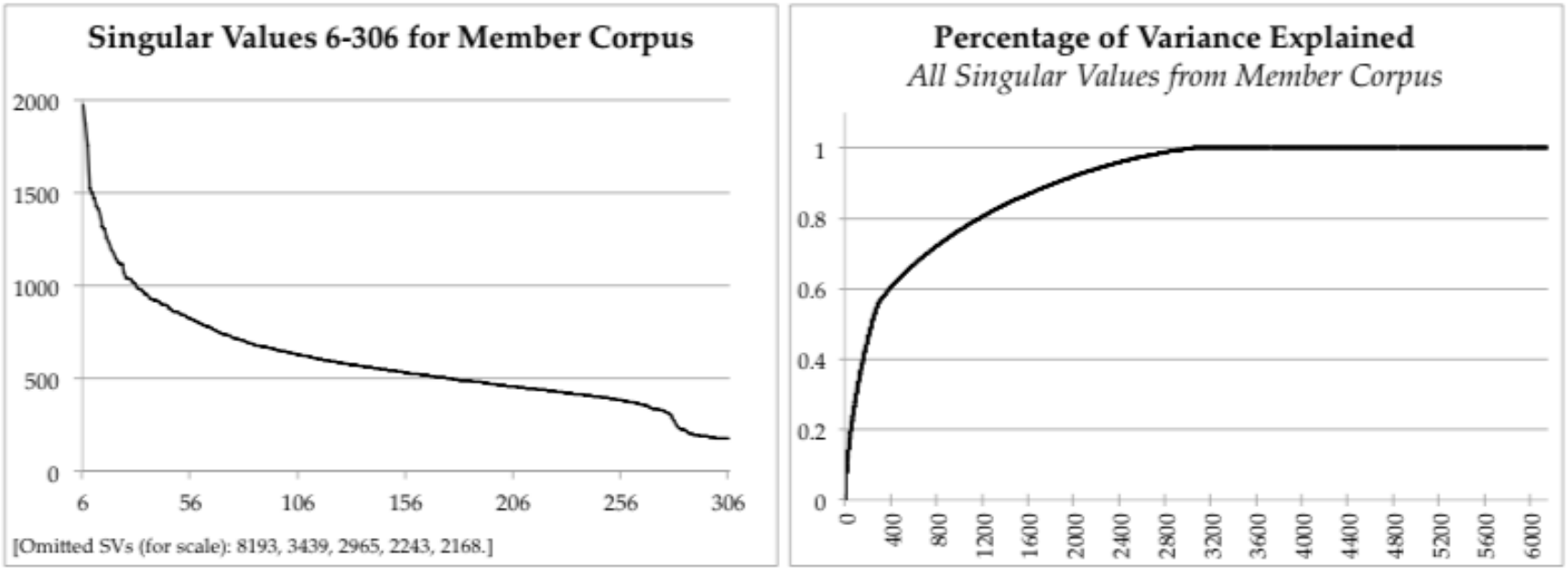
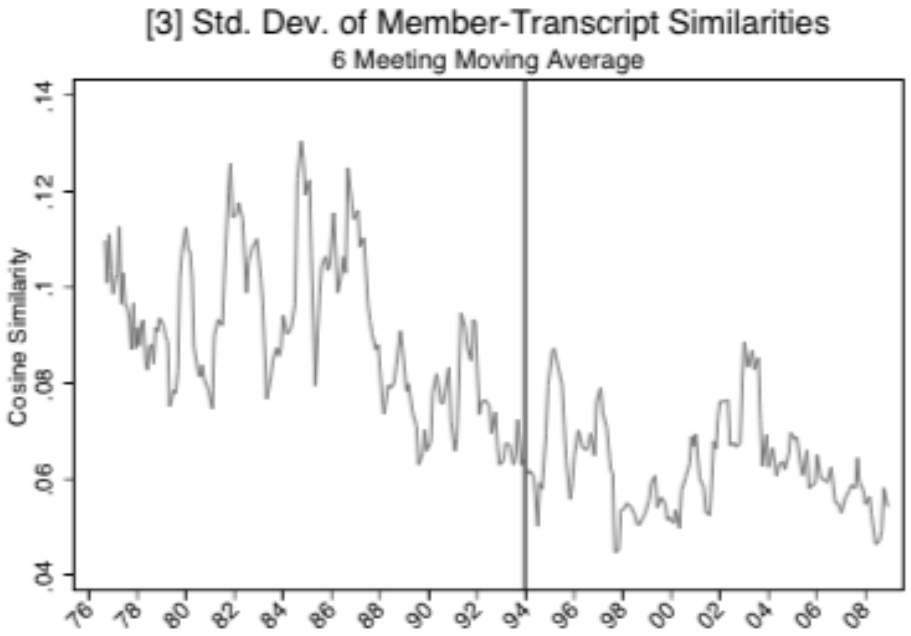
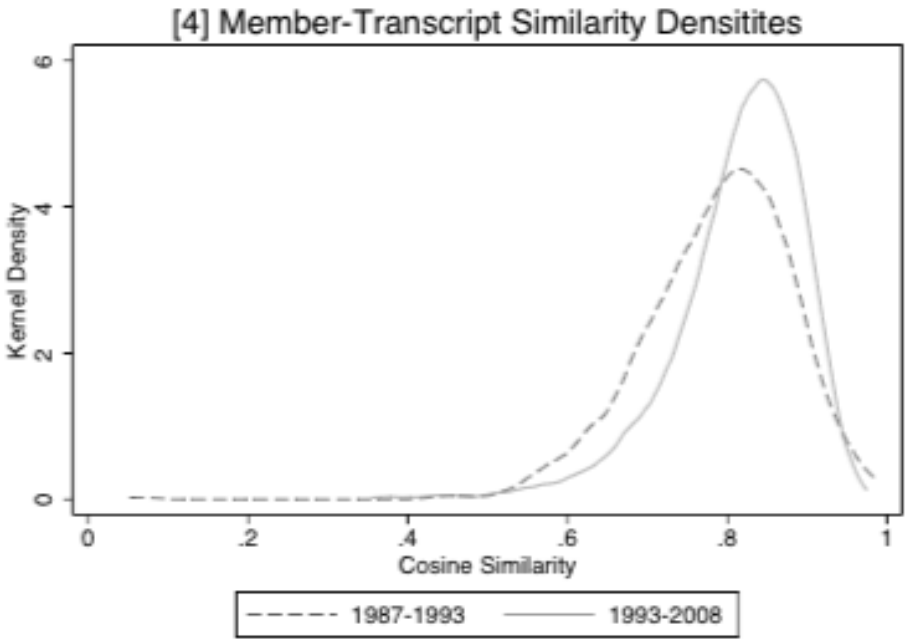


Figure 11: The left hand side shows the 6th through 306th singular values (the elements $\sigma_i \in \mathbf{\Sigma}$ from the SVD) from the member corpus. The right hand side graph show percentage of the variance explained by all 6152 singular values for the member corpus.

Results


[3] Std. Dev. of Member-Transcript Similarities
6 Meeting Moving Average
Cosine Similarity
.04
.06
.08
.1
.12
.14
76
78
80
82
84
86
88
90
92
94
96
98
00
02
04
06
08


[4] Member-Transcript Similarity Densitites
Kernel Density
0
2
4
6
0
.2
.4
.6
.8
1
Cosine Similarity
1987-1993
1993-2008

Interpretation

Finding are consistent with a tendency towards less disagreement with more transparency.

This is consistent with economic theory, and with previous work by Meade and Stasavage on FOMC voting behavior.

More on this later.

Limitations of LSA

LSA is an important first step in thinking about content in terms of latent variables.

But there are some limitations:

1. Hard to interpret singular vectors as topics—they are lines in $\mathbb{R}^k$.
2. Deterministic mapping between topics and words/documents.
3. Frobenius norm minimization appropriate for normal random variables, but elements of $\mathbf{X}$ are discrete (and tf-idf values lie in $\mathbb{R}^+$).

See also work of Schonhardt-Bailey.