Machine Learning Methods for Economists
Unsupervised Learning

Stephen Hansen
University of Oxford

Introduction

Recall that unsupervised learning seeks to uncover hidden structure in observations.

There may be several motivations for this:

1. Describe the most prominent sources of variation within a vast array of covariates.
2. Find a low-dimensional representation of a high-dimensional object that preserves most relevant information.
3. Group observations according to similarity.

Unsupervised learning algorithms already popular in economics: principal components, factor models, clustering algorithms.

We introduce some simple non-parametric algorithms for discrete data before introducing probabilistic structure that builds on previous slides.

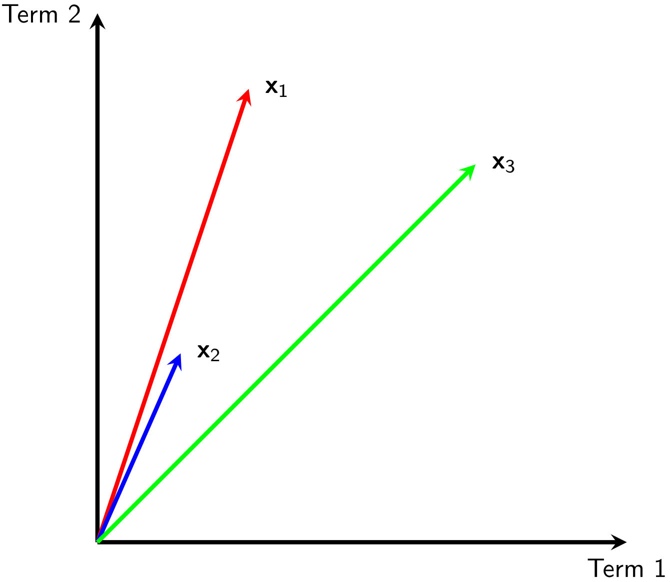
Vector Space Model

One can view the rows of the document-term matrix as vectors lying in a V-dimensional space.

The basis for the vector space is $e_1, \dots, e_V$.

The question of interest is how to measure the similarity of two documents in the vector space, and whether unsupervised learning can help with this.

Three Documents



Cosine Similarity

Define the cosine similarity between documents i and j as

$$CS(i,j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|}$$

1. Since document vectors have no negative elements $CS(i,j) \in [0,1]$.
2. $\mathbf{x}_i / \|\mathbf{x}_i\|$ is unit-length, correction for different distances.

Information Retrieval

The problem of *synonomy* is that several different words can be associated with the same topic. Cosine similarity between following documents?

school	university	college	teacher	professor
0	5	5	0	2

school	university	college	teacher	professor
10	0	0	4	0

The problem of *polysemy* is that the same word can have multiple meanings. Cosine similarity between following documents?

tank	seal	frog	animal	navy	war
10	10	3	2	0	0

tank	seal	frog	animal	navy	war
10	10	0	0	4	3

If we correctly map words into topics, comparisons become more accurate.

Outline

	Non-parametric	Parametric
1d latent representation	1. k-means	3. mixture model
$>$1d latent representation	2. PCA/LSI	4. pLSI/LDA

K-Means

Recall we can represent document d as a vector $\vec{x}_d \in \mathbb{R}_+^V$. In the k-means model, every document has a single cluster assignment.

Let D_k be the set of all documents that are in cluster k. The *centroid* of the documents in cluster k is $\vec{u}_k = \frac{1}{|D_k|} \sum_{d \in D_k} \vec{x}_d$.

In k-means we choose cluster assignments $\{D_1, \ldots, D_K\}$ to minimize the sum of squares between each document and its cluster centroid:

$$\sum_k \sum_{d \in D_k} \|\vec{x}_d - \vec{u}_k\|^2$$

Solution groups similar documents together, and centroids represent prototype documents within each cluster.

Normalize document lengths to cluster on content, not length.

Solution Algorithm

First initialize the centroids $\vec{u}_k$ for $1, \ldots, K$.

Repeat the following steps until convergence:

1. Assign each document to its closest centroid, i.e. choose an assignment k for d that minimizes $\|\vec{x}_d - \vec{u}_k\|$.
2. Recompute the cluster centroids as $\vec{u}_k = \frac{1}{|D_k|} \sum_{d \in D_k} \vec{x}_d$ given the updated assignments in previous step.

The objective function is guaranteed to decrease at each step $\rightarrow$ convergence to local minimum.

Proof: for step 1 obvious; for step 2 choose elements of vector $\vec{y} \in \mathbb{R}_+^V$ to minimize $\sum_{d \in D_k} \|\vec{x}_d - \vec{y}\|^2 \equiv \sum_{d \in D_k} \sum_v (x_{d,v} - y_v)^2$. Solution is exactly $\vec{u}_k$.

Mixed-Membership Models

In k-means, documents are associated with a single topic.

In practice, we might imagine that documents cover more than one topic.

Examples: State-of-the-Union Addresses discuss domestic <u>and</u> foreign policy; monetary policy speeches discuss inflation <u>and</u> growth.

Models that associated observations with more than one latent variable are called *mixed-membership* models. Also relevant outside of text mining: in models of group formation, agents can be associated with different latent communities (sports team, workplace, church, etc).

Latent Semantic Analysis

One of the first mixed-membership models in text mining was the Latent Semantic Analysis/Indexing model of Deerwester et. al. (1990).[1]

A linear algebra approach that applies a singular value decomposition to document-term matrix.

Closely related to classical principal components analysis.

Examples in economics: Boukus and Rosenberg (2006);[2] Hendry and Madeley (2010);[3] Acosta (2014);[4] Waldinger et. al. (2018).[5]

[1] Journal of the American Society for Information Science, "Indexing by Latent Semantic Analysis".

[2] WP, "The Information Content of FOMC Minutes".

[3] WP, "Text Mining and the Information Content of Bank of Canada Communications".

[4] WP, "FOMC Responses to Calls for Transparency".

[5] QJE, "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science"

Review

Let $\mathbf{X}$ be an $N \times N$ symmetric matrix with N linearly independent eigenvectors.

Then there exists a decomposition $\mathbf{X} = \mathbf{Q \Lambda Q}^T$, where $\mathbf{Q}$ is an orthogonal matrix whose columns are eigenvectors of $\mathbf{X}$ and $\mathbf{\Lambda}$ is a diagonal matrix whose entries are eigenvalues of $\mathbf{X}$.

When we apply this decomposition to the variance-covariance matrix of a dataset, we can perform principal components analysis.

The eigenvalues in $\mathbf{\Lambda}$ give a ranking of the columns in $\mathbf{Q}$ according to the variance they explain in the data.

Singular Value Decomposition

The document-term matrix $\mathbf{X}$ is not square, but we can decompose it using a generalization of the eigenvector decomposition called the *singular value decomposition.*

Proposition

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

Singular Value Decomposition

The document-term matrix $\mathbf{X}$ is not square, but we can decompose it using a generalization of the eigenvector decomposition called the *singular value decomposition.*

Proposition

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

Interpretation of Left Singular Vectors

Note that $\mathbf{X}\mathbf{X}^T = \mathbf{A}\mathbf{\Sigma}\mathbf{B}^T\mathbf{B}\mathbf{\Sigma}^T\mathbf{A}^T = \mathbf{A}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{A}^T$.

This is the eigenvector decomposition of the matrix $\mathbf{X}\mathbf{X}^T$, whose (i,j)th element measures the overlap between documents i and j.

Left singular vectors are eigenvectors of $\mathbf{X}\mathbf{X}^T$ and σ_i^2 are associated eigenvalues.

Interpretation of Right Singular Vectors

Note that $\mathbf{X}^T\mathbf{X} = \mathbf{B}\mathbf{\Sigma}^T\mathbf{A}^T\mathbf{A}\mathbf{\Sigma}\mathbf{B}^T = \mathbf{B}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{B}^T$.

This is the eigenvector decomposition of the matrix $\mathbf{X}^T\mathbf{X}$, whose (i, j)th element measures the overlap between terms i and j.

Right singular vectors are eigenvectors of $\mathbf{X}^T\mathbf{X}$ and σ_i^2 are associated eigenvalues.

Approximating the Document-Term Matrix

We can obtain a rank k approximation of the document-term matrix $\mathbf{X}_k$ by constructing $\mathbf{X}_k = \mathbf{A}\mathbf{\Sigma}_k\mathbf{B}^T$, where $\mathbf{\Sigma}_k$ is the diagonal matrix formed by replacing $\mathbf{\Sigma}_{ii} = 0$ for $i > k$.

The idea is to keep the "content" dimensions that explain common variation across terms and documents and drop "noise" dimensions that represent idiosyncratic variation.

Often k is selected to explain a fixed portion p of variance in the data. In this case k is the smallest value that satisfies $\sum_{i=1}^{k} \sigma_i^2 / \sum_i \sigma_i^2 \geq p$.

We can then perform the same operations on $\mathbf{X}_k$ as on $\mathbf{X}$, e.g. cosine similarity.

Example

Suppose the document-term matrix is given by

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

Matrix of Cosine Similarities

$$
\begin{array}{c c}
 & \begin{array}{cccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \end{array} \\
\begin{array}{c} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{array} &
\begin{bmatrix}
1 & \cdot & \cdot & \cdot & \cdot & \cdot \\
0.70 & 1 & \cdot & \cdot & \cdot & \cdot \\
0.00 & 0.69 & 1 & \cdot & \cdot & \cdot \\
0.08 & 0.30 & 0.17 & 1 & \cdot & \cdot \\
0.10 & 0.21 & 0.00 & 0.92 & 1 & \cdot \\
0.02 & 0.17 & 0.00 & 0.66 & 0.88 & 1
\end{bmatrix}
\end{array}
$$

SVD

The singular values are $(31.61, 15.14, 10.90, 5.03)$.

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

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

Rank-2 Approximation

$$\mathbf{X}_2 = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{array} \begin{array}{c} \begin{array}{cccc} \text{car} & \text{automobile} & \text{ship} & \text{boat} \end{array} \\ \begin{bmatrix} 0.5343 & 2.1632 & 0.8378 & 0.7169 \\ 1.3765 & 5.8077 & 1.2765 & 0.9399 \\ 2.9969 & 13.2992 & 0.3153 & 0.4877 \\ 0.8817 & 1.9509 & 7.4715 & 7.4456 \\ 1.1978 & 0.0670 & 20.3682 & 20.6246 \\ 0.2219 & 0.1988 & 4.4748 & 4.5423 \end{bmatrix} \end{array}$$

Matrix of Cosine Similarities

$$
\begin{array}{c c}
 & \begin{array}{cccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \end{array} \\
\begin{array}{c} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{array} &
\begin{bmatrix}
1 & \cdot & \cdot & \cdot & \cdot & \cdot \\
0.97 & 1 & \cdot & \cdot & \cdot & \cdot \\
0.91 & 0.97 & 1 & \cdot & \cdot & \cdot \\
0.60 & 0.43 & 0.23 & 1 & \cdot & \cdot \\
0.45 & 0.26 & 0.05 & 0.98 & 1 & \cdot \\
0.47 & 0.29 & 0.07 & 0.98 & 0.99 & 1
\end{bmatrix}
\end{array}
$$

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

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 econ-

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

Deal Signed in Ukraine, but Shows St

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

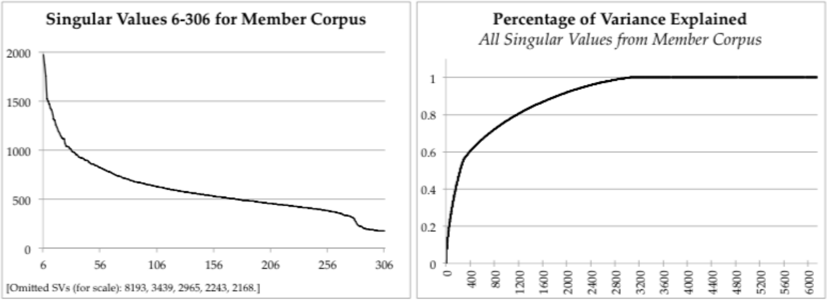
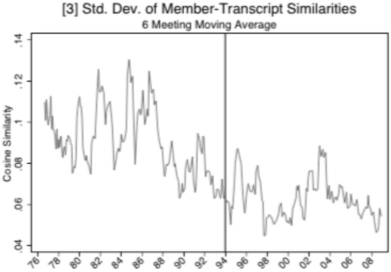
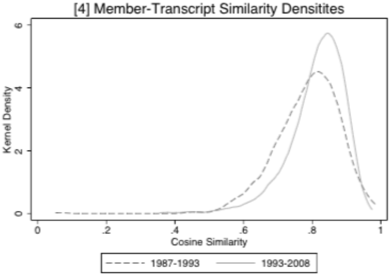


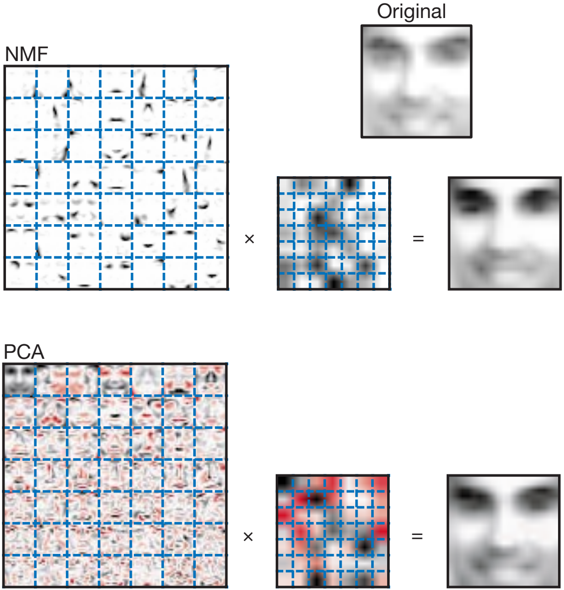
Figure 11: The left hand side shows the 6th through 306th singular values (the elements $\sigma_i \in \Sigma$ from the SVD) from the member corpus. The right hand side graph show percentage of the variance explained by all 6152 singular values for the member corpus.

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

Non-negative Matrix Factorization[6]



[6] Lee and Seung (1999, Nature) "Learning the parts of objects by non-negative matrix factorization"

Probabilistic Modeling

The k-means and LSI models are useful tools for data exploration, but have no immediately obvious statistical foundations relevant to text.

A more satisfactory approach might be to write down a statistical model for documents whose parameters we estimate—allows us to incorporate and make inferences about relevant structure in the corpus.

We can draw on our discussion of probability models in the previous lecture to build a generative latent variable model.

Generative LVM

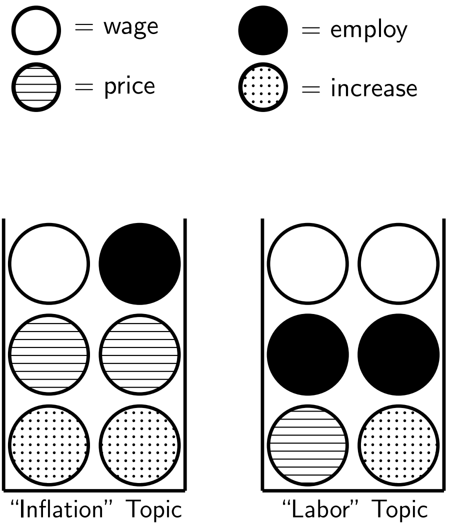
All of the generative latent variable models we will discuss have the form $\mathbf{x}_d \sim \text{MN}(\sum_k \theta_{d,k}\boldsymbol{\beta}_k, N_d)$.

This builds on the simple language model from the previous lecture, but introduces k separate categorical distributions, each with parameter vector $\boldsymbol{\beta}_k$. The probability that topic k generate term v is $\beta_{k,v}$.

Each document is represented on a space of topics with $\boldsymbol{\theta}_d \in \Delta^{K-1}$ instead of a raw vocabulary space as with $\mathbf{x}_d \in \mathbb{Z}_+^V$. $\theta_{d,k}$ is the share of topic k in document d.

Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D)$.

Topics as Urns



Multinomial Mixture Model

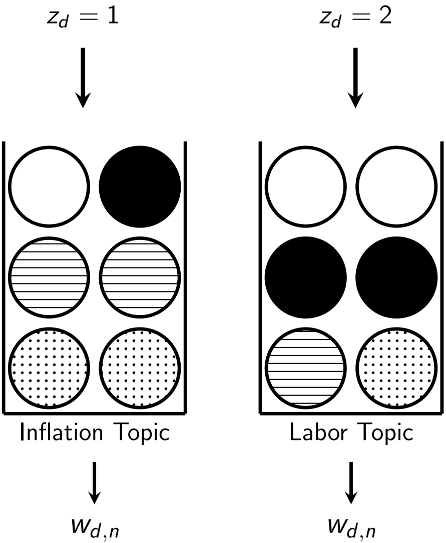
Latent variable models differ in how they model $\boldsymbol{\theta}_d = (\theta_{d,1}, \ldots, \theta_{d,K})$.

In a *mixture model*, every document belongs to a single category $z_d \in \{1, \ldots, K\}$, which is independent across documents and drawn from $\Pr[\, z_d = k \,] = \rho_k$.

We then have

$$\theta_{d,k} = \begin{cases} 1 & \text{if } z_d = k \\ 0 & \text{otherwise} \end{cases}.$$

Mixture Model for Document



Probability of Document

We can derive the probability of observing a document given the vector of mixing probabilities $\boldsymbol{\rho}$ and the matrix of term probabilities $\mathbf{B}$.

Suppose that $z_d = k$. Then the probability of $\mathbf{w}_d$ is $\prod_v (\beta_{k,v})^{x_{d,v}}$.

To compute the unconditional probability of document d, we need to marginalize over the latent assignment variable z_d

$$\Pr[\mathbf{x}_d \mid \boldsymbol{\rho}, \boldsymbol{\beta}] = \sum_{z_d} \Pr[\mathbf{x}_d \mid z_d, \boldsymbol{\rho}, \boldsymbol{\beta}] \Pr[z_d \mid \boldsymbol{\rho}, \boldsymbol{\beta}] = \sum_k \rho_k \prod_v (\beta_{k,v})^{x_{d,v}}.$$

Probability of Corpus

By independence of latent variables across documents, the likelihood of entire corpus, which we can summarize with document-term matrix $\mathbf{X}$ is

$$L(\mathbf{X} \mid \boldsymbol{\rho}, \boldsymbol{\beta}) = \prod_d \sum_k \rho_k \prod_v (\beta_{k,v})^{x_{d,v}}$$

and log-likelihood is

$$\ell(\mathbf{X} \mid \boldsymbol{\rho}, \boldsymbol{\beta}) = \sum_d \log \left(\sum_k \rho_k \prod_v (\beta_{k,v})^{x_{d,v}} \right).$$

Inference

Here the sum over the latent variable assignments lies within the logarithm, which makes MLE intractable.

On the other hand, if we knew the category assignment of each document, MLE would be very easy.

We can therefore use the expectation-maximization (EM) algorithm for parameter inference.

EM Algorithm

First initialize parameter values $\boldsymbol{\rho}^0$ and $\boldsymbol{\beta}^0$. Then, at iteration i:

1. (E-step). Compute the posterior distribution over the latent variables $\mathbf{z}_d = (z_1, \ldots, z_D)$ given $\boldsymbol{\rho}^{i-1}, \boldsymbol{\beta}^{i-1}$ and data. Use this distribution to form

$$Q(\boldsymbol{\rho}, \boldsymbol{\beta}, \boldsymbol{\rho}^{i-1}, \boldsymbol{\beta}^{i-1}) \equiv \mathbb{E}_z[\,\ell_{\text{comp}}\,(\mathbf{X}, \mathbf{z} \mid \boldsymbol{\rho}, \boldsymbol{\beta})\,].$$

2. (M-step). Update parameter estimates to $\boldsymbol{\rho}^i, \boldsymbol{\beta}^i$ by maximizing $Q(\boldsymbol{\rho}, \boldsymbol{\beta}, \boldsymbol{\rho}^{i-1}, \boldsymbol{\beta}^{i-1})$ with respect to $\boldsymbol{\rho}, \boldsymbol{\beta}$.

3. If convergence criterion met, stop; otherwise proceed to iteration $i = i + 1$.

The log-likelihood $\ell\,(\mathbf{X} \mid \boldsymbol{\rho}, \boldsymbol{\beta})$ is guaranteed to increase at each iteration. We converge to a local maximum.

Complete Data Log-Likelihood

The joint distribution of $\mathbf{x}_d$ and z_d is $\prod_k \left[\rho_k \prod_v (\beta_{k,v})^{x_{d,v}}\right]^{\mathbb{1}(z_d=k)}$ and so the joint distribution of $\mathbf{X}$ and $\mathbf{z} = (z_1, \ldots, z_D)$ is

$$L_{\mathrm{comp}}\left(\mathbf{X}, \mathbf{z} \mid \boldsymbol{\rho}, \boldsymbol{\beta}\right) = \prod_d \prod_k \left[\rho_k \prod_v (\beta_{k,v})^{x_{d,v}}\right]^{\mathbb{1}(z_d=k)}$$

The *complete data log-likelihood* is

$$\ell_{\mathrm{comp}}\left(\mathbf{X}, \mathbf{z} \mid \boldsymbol{\rho}, \boldsymbol{\beta}\right) = \sum_d \sum_k \mathbb{1}(z_d = k) \left[\log(\rho_k) + \sum_v x_{d,v} \log\left(\beta_{k,v}\right)\right].$$

Note that this function is much easier to maximize with respect to the parameters than the original log-likelihood function.

Expectation Step

Compute expected value of the complete data log-likelihood with respect to the latent variables given the current value of the parameters $\boldsymbol{\rho}^i$ and $\boldsymbol{\beta}^i$ and data.

Expectation Step

Compute expected value of the complete data log-likelihood with respect to the latent variables given the current value of the parameters $\boldsymbol{\rho}^i$ and $\boldsymbol{\beta}^i$ and data.

Clearly $\mathbb{E}\left[\mathbb{1}(z_d = k) \,\middle|\, \boldsymbol{\rho}^i, \boldsymbol{\beta}^i, \mathbf{X}\right] = \Pr\left[z_d = k \,\middle|\, \boldsymbol{\rho}^i, \boldsymbol{\beta}^i, \mathbf{X}\right] \equiv \widehat{z}_{d,k}$.

By Bayes' Rule we have that

$$
\begin{aligned}
\widehat{z}_{d,k} = &\Pr\left[z_d = k \,\middle|\, \boldsymbol{\rho}^i, \boldsymbol{\beta}^i, \mathbf{x}_d\right] \propto \\
&\Pr\left[\mathbf{x}_d \,\middle|\, \boldsymbol{\rho}^i, \boldsymbol{\beta}^i, z_d = k\right] \Pr\left[z_d = k \,\middle|\, \boldsymbol{\rho}^i, \boldsymbol{\beta}^i\right] = \rho_k \prod_v (\beta_{k,v})^{x_{d,v}}.
\end{aligned}
$$

So the expected complete log-likelihood becomes

$$
Q(\boldsymbol{\rho}, \boldsymbol{\beta}, \boldsymbol{\rho}^i, \boldsymbol{\beta}^i) = \sum_d \sum_k \widehat{z}_{d,k} \left[\log(\rho_k) + \sum_v x_{d,v} \log(\beta_{k,v})\right]
$$

Maximization Step

Maximize the expected complete log-likelihood with respect to $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$.

Maximization Step

Maximize the expected complete log-likelihood with respect to $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$. The Lagrangian for this problem is

$$Q(\boldsymbol{\rho}, \boldsymbol{\beta}, \boldsymbol{\rho}^i, \boldsymbol{\beta}^i) + \nu \left(1 - \sum_k \rho_k\right) + \sum_k \lambda_k \left(1 - \sum_v \beta_{k,v}\right).$$

Maximization Step

Maximize the expected complete log-likelihood with respect to $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$. The Lagrangian for this problem is

$$Q(\boldsymbol{\rho}, \boldsymbol{\beta}, \boldsymbol{\rho}^i, \boldsymbol{\beta}^i) + \nu \left(1 - \sum_k \rho_k\right) + \sum_k \lambda_k \left(1 - \sum_v \beta_{k,v}\right).$$

Standard maximization gives

$$\rho_k^{i+1} = \frac{\sum_d \hat{z}_{d,k}}{\sum_k \sum_d \hat{z}_{d,k}},$$

or the average probability that documents have topic k and

$$\beta_{k,v}^{i+1} = \frac{\sum_d \hat{z}_{d,k} x_{d,v}}{\sum_d \hat{z}_{d,k} \sum_v x_{d,v}},$$

or the expected number of times documents of type k generate term v over the expected number of words generated by type k documents.

Example

Let $K = 2$ and consider the corpus of 1,232 paragraphs of State-of-the-Union Addresses since 1900.

Topic	Top Terms
0	tax.job.help.must.congress.need.health.care.busi.let.school.time
1	world.countri.secur.must.terrorist.iraq.state.energi.help.unit

$(\rho_0, \rho_1) = (0.42, 0.58)$.

No *ex ante* labels on clusters, so any interpretation is *ex post*, and potentially subjective, judgment on the part of the researcher.

K-Means as EM

The k-means algorithm can be viewed as the EM algorithm under a special case of a Gaussian mixture model in which the distribution of data in cluster k is $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ where $\Sigma_k = \sigma^2 \mathbf{I}_V$ and σ^2 is small.

The probability that document d is generated by the cluster with the closest mean is then close to 1, so the assignment of documents to the closest centroid in the k-means algorithm is the E-step.

Given the spherical covariance matrix, the probability of observing documents within cluster k is proportional to the sum of squared distances between documents and the mean. So recomputing the cluster centroids is the M-step.

Good news is that k-means has statistical foundations; bad news is that the appropriateness of these for count data is doubtful.

Probabilistic Mixed-Membership Model

As with k-means, LSA provides a useful tool for data exploration, but its statistical foundations are unclear.

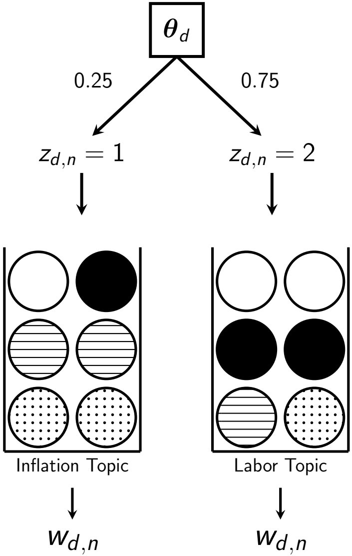
Recall the generative model $\mathbf{x}_d \sim \text{MN}(\sum_k \theta_{d,k} \boldsymbol{\beta}_k, N_d)$.

In the probabilistic LSA model of Hofmann (1999) we allow $\boldsymbol{\theta}$ to lie anywhere in the $K-1$ simplex.

Instead of assigning each document to a topic, we can assign each <u>word</u> in each document to a topic.

Let $z_{d,n} \in \{1, \ldots, K\}$ be the topic assignment of $w_{d,n}$;
$\mathbf{z}_d = (z_{d,1}, \ldots, z_{d,N_d}$; and $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_D)$.

Mixed-Membership Model for Document



Estimation

The likelihood function for this model is

$$\prod_d \prod_n \sum_{z_{d,n}} \Pr\left[w_{d,n} \;\middle|\; \boldsymbol{\beta}_{z_{d,n}} \right] \Pr\left[z_{d,n} \mid \boldsymbol{\theta}_d \right].$$

We can fit the parameters by EM, but:

1. Large number of parameters $KV + DK$, prone to over-fitting.
2. No generative model for $\boldsymbol{\theta}_d$.

We will come back to these issues in the next lecture.

Word Embeddings

In machine learning and natural language processing, word embeddings are currently nearer the frontier for dimensionality reduction; word2vec is a particularly popular algorithm.

The idea is to construct a low-dimensional representation for each vocabulary term in the corpus by explicitly modeling the probability of seeing each word given a “context” of surrounding words.

The resulting embedding vectors capture more semantic meaning than LSA.

For example, word2vec can perform analogy tasks: vector for ‘paris’ minus ‘france’ plus ‘italy’ is close to ‘rome’.

Exponential Family Embeddings

The embeddings idea has been extended by Rudolph et. al. (2016),[7] and applied to shopping basket data.

Ruiz et. al. (2016)[8] extends the idea to incorporate prices and sequential choice.

The model provides a flexible way of identifying complements and substitutes based on the co-occurrence patterns of items in shopping baskets:

query items	complementarity score		exchangeability score	
mission tortilla soft taco	2.51	ortega taco shells white corn	0.05	mission fajita size
	2.40	mcrmck seasoning mix taco	0.10	mission tortilla fluffy gordita
	2.26	lawrys taco seasoning mix	0.11	mission tortilla soft taco
private brand hot dog buns	3.02	bp franks bun size	0.10	private brand hamburger buns
	2.94	bp franks beef bun length	0.12	ball park buns hot dog
	2.86	private brand hamburger buns	0.14	private brand hot dog buns ssme 8ct

[7] NIPS, "Exponential Family Embeddings"

[8] WP, "SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements".

Conclusion

Key ideas from this lecture:

1. The goal of unsupervised learning is to estimate latent structure in observations.
2. Mixture versus mixed-membership models.
3. Ad hoc data exploration tools are a good starting point, but probabilistic models are more flexible and statistically well-founded.
4. EM algorithm for likelihood functions with latent variables.