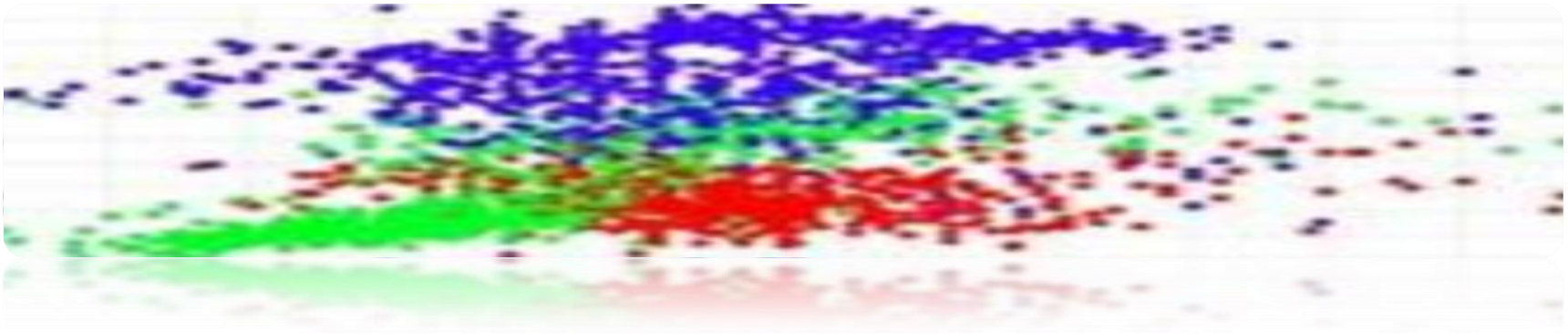
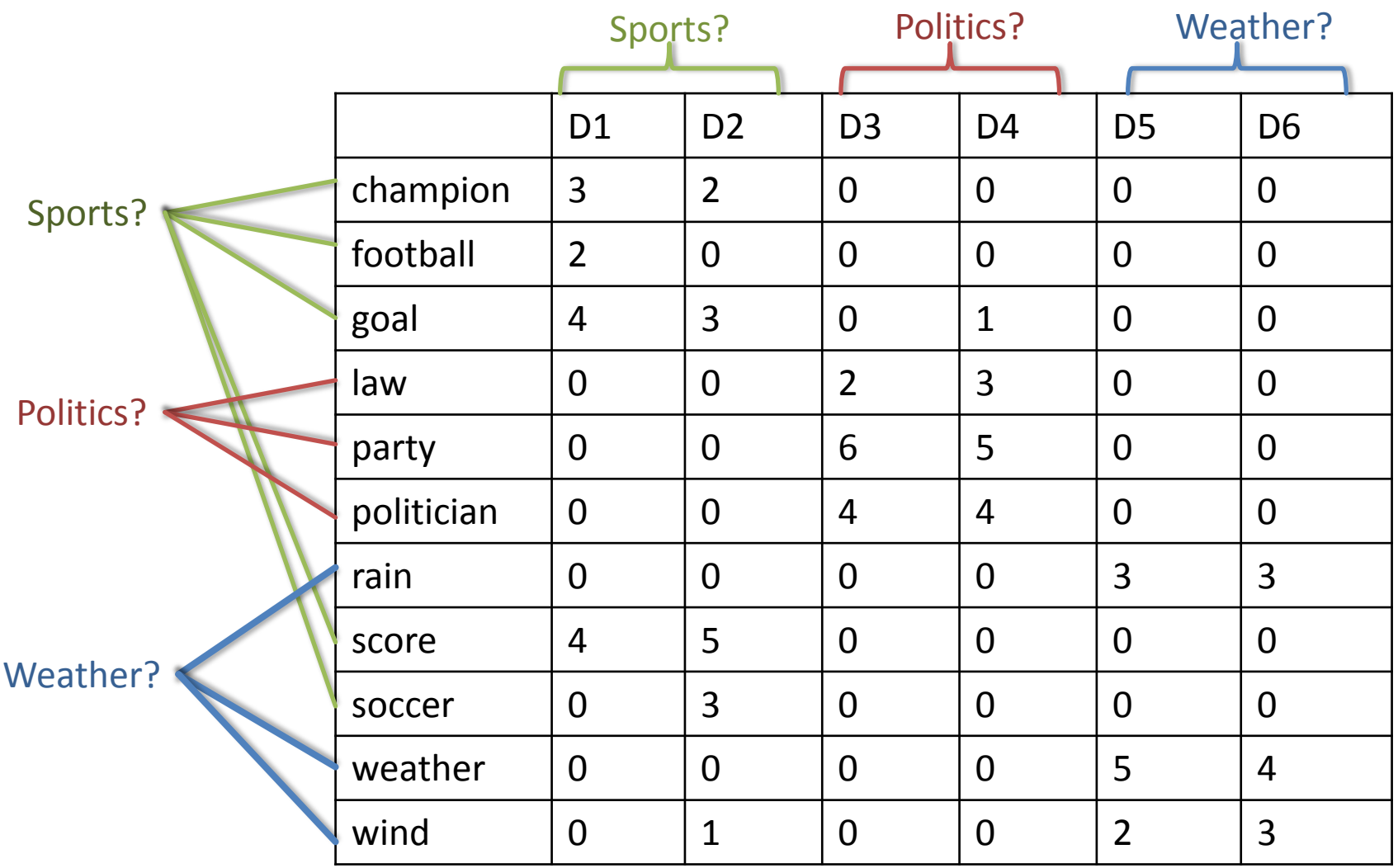


UNSUPERVISED LEARNING – TOPIC MODELS

Outline

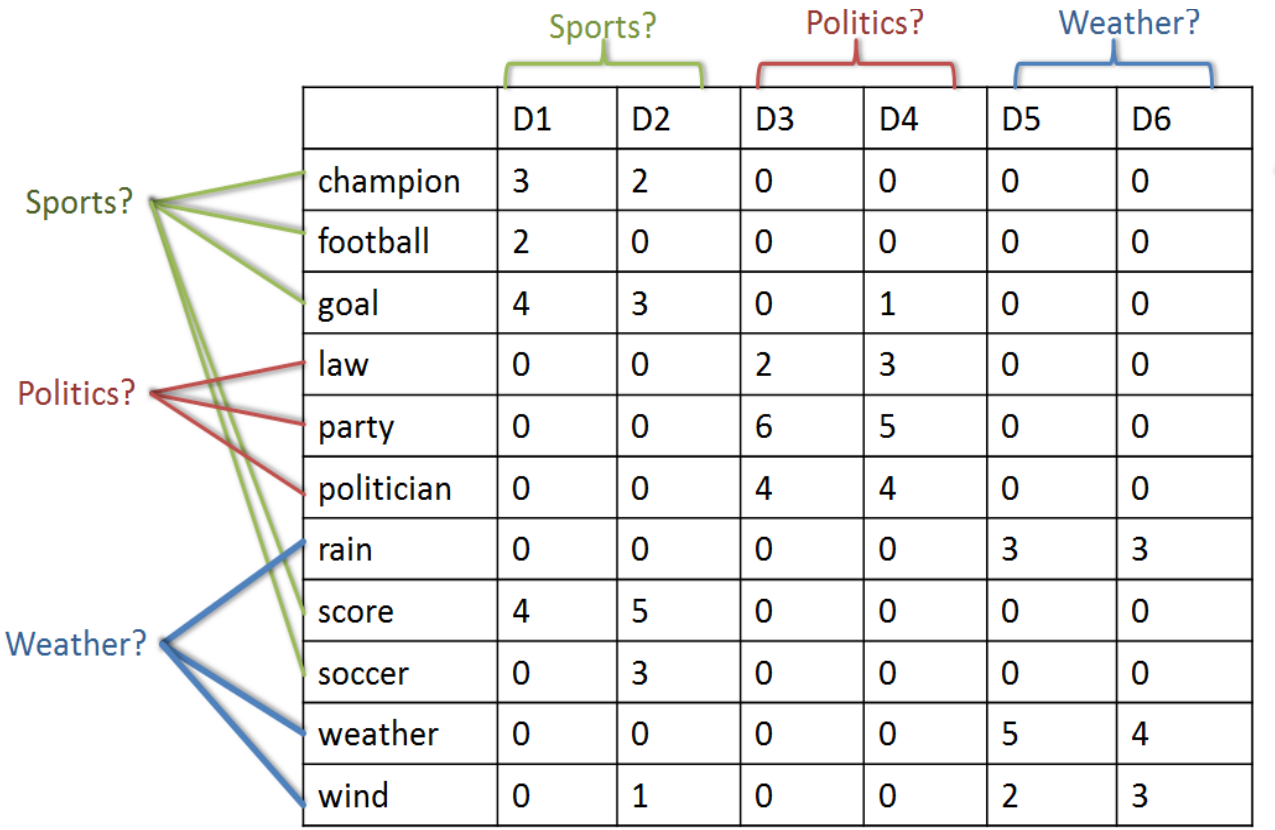
- Latent topics
- Basics of Linear Algebra
- Latent Semantic Analysis
- Probabilistic Latent Semantic Analysis
- Latent Dirichlet Allocation

Latent topics (1)



	D1	D2	D3	D4	D5	D6
champion	3	2	0	0	0	0
football	2	0	0	0	0	0
goal	4	3	0	1	0	0
law	0	0	2	3	0	0
party	0	0	6	5	0	0
politician	0	0	4	4	0	0
rain	0	0	0	0	3	3
score	4	5	0	0	0	0
soccer	0	3	0	0	0	0
weather	0	0	0	0	5	4
wind	0	1	0	0	2	3

Latent topics (2)



	D1	D2	D3	D4	D5	D6
champion	3	2	0	0	0	0
football	2	0	0	0	0	0
goal	4	3	0	1	0	0
law	0	0	2	3	0	0
party	0	0	6	5	0	0
politician	0	0	4	4	0	0
rain	0	0	0	0	3	3
score	4	5	0	0	0	0
soccer	0	3	0	0	0	0
weather	0	0	0	0	5	4
wind	0	1	0	0	2	3

Typically very sparse

If two documents have high overlap of words they are probably about the same topic

Can we model this?

Possible approaches to topic models

- Algebraic approach
  - Given: $m \times n$ word-document matrix (m words, n documents)
  - Question: Is it possible to map documents from high-dim. (i.e., m-dim. vector space into k-dim. vector space, with $k \ll m$ and preserve, or even reinforce their similarity?
- Probabilistic approach
  - Given: $m \times n$ word-document matrix (m words, n documents)
  - Question: Is it possible to estimate the joint probability of terms and docs as a decomposition of term-doc distributions over $k \ll m$ topics?

Basics: Linear independence, rank, and basis

- Vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are **linearly independent** if $\sum_{i=1}^{k} \alpha_i \mathbf{v}_i = 0$ is only possible for $\alpha_i = 0, \forall\, i$
- The **rank** of a matrix is the maximal number of independent row or column vectors
- The **basis** of an $n \times n$ matrix is the set *S* of row (column) vectors such that all rows (columns) are linear combinations of vectors from *S*
- *S* is an **orthonormal basis** if for all $\mathbf{v}_i, \mathbf{v}_j \in S$: $\|\mathbf{v}_i\|_2 = 1 = \|\mathbf{v}_j\|_2$ and $\mathbf{v}_i{}^T \mathbf{v}_j = 0$

Basics: Eigenvectors and Eigenvalues

- For an $n \times n$ matrix A, $n \times 1$ vector $\mathbf{v}$ and a scalar λ that satisfy $A\mathbf{v} = \lambda\mathbf{v}$ are called **Eigenvector** and **Eigenvalue** of A

- Eigenvalues are the roots of the **characteristic function**

$$f(\lambda) = \det(A - \lambda I), \text{ with}$$

$$\det(A) = \sum_{\substack{i=1 \\ j=1}}^{i=n} (-1)^{(i+j)} a_{ij} \det(A^{(\backslash \mathrm{ij})})$$

$A^{(\backslash \mathrm{ij})}$ is A without the i'th row and the j'th column

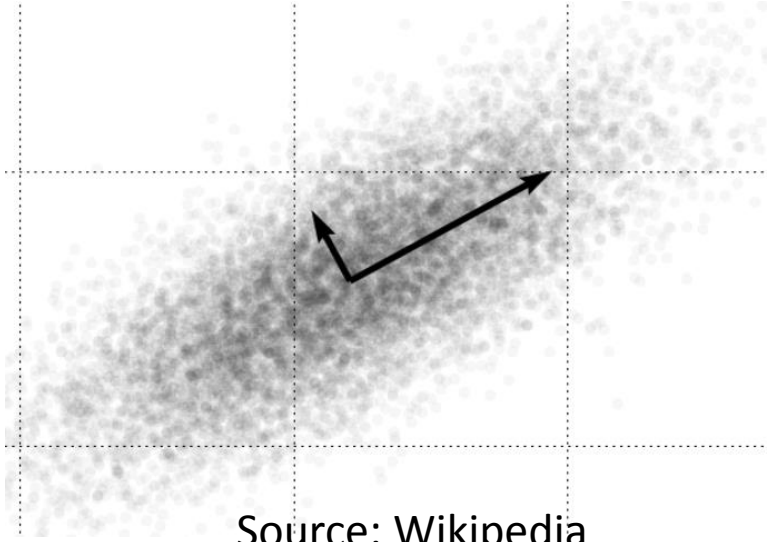
- Furthermore, if A is symmetric, all Eigenvalues are real!

Spectral (or principal components) theorem

- For an $n \times n$ symmetric matrix A with Eigenvalues $\lambda_1, \dots, \lambda_n$ and Eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ forming an orthonormal basis of A ($\|\mathbf{v}_i\|_2 = 1, \forall i$) the following holds:

$$D = (\mathbf{v}_1 \dots \mathbf{v}_n)^T A (\mathbf{v}_1 \dots \mathbf{v}_n) \text{ and } A = (\mathbf{v}_1 \dots \mathbf{v}_n) D (\mathbf{v}_1 \dots \mathbf{v}_n)^T,$$

where D is the diagonal matrix holding $\lambda_1, \dots, \lambda_n$



Source: Wikipedia

- For orthogonal Eigenvectors, each Eigenvector represents the variability of the data in one dimension!
- If Eigenvectors form orthonormal basis, they have equal length their importance is encoded in the corresponding Eigenvalues!
- Can we generalize this for $m \times n$ matrices?

Singular value decomposition (SVD) theorem

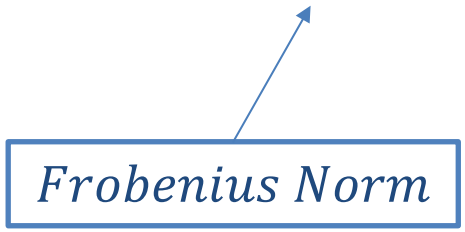
- For every real-valued $m \times n$ matrix A with rank r, there exists a decomposition (i.e., factorization) $A = UDV^T$ with
  - U: $m \times r$ matrix with orthonormal column vectors
  - D: $r \times r$ diagonal matrix
  - V: $n \times r$ matrix with orthonormal column vectors

- The SVD is unique if the elements of D, i.e. the **singular values** of A, are sorted

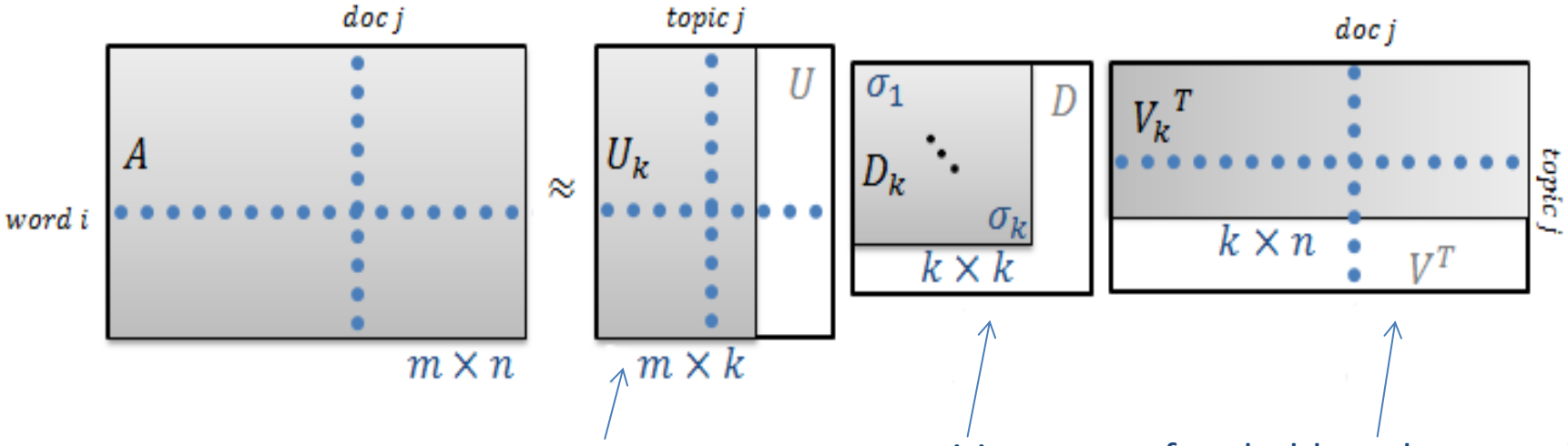
- Based on the **Spectral Theorem** the SVD of a real-valued $m \times n$ matrix A (i.e., $A = UDV^T$) can be achieved as follows:
  - D contains the positive roots of the Eigenvalues of $A^T A$, i.e. singular values of A, (these exist because $A^T A$ is **symmetric** and **positive definite:** $\mathbf{v}^T A^T A \mathbf{v} > 0, \forall\, \mathbf{v} \neq \vec{0}$)
  - U contains the Eigenvectors of AA^T as columns
  - V contains the Eigenvectors of $A^T A$ as columns

Low-rank approximation theorem

- For an $m \times n$ matrix A with rank r, let $A' := U_k D_k {V_k}^T$ be an approximation of A, with $k \leq r$, D_k containing the top-k largest values of D, U_k and V_k containing the corresponding Eigenvectors. Among all $m \times n$ matrices M, A' minimizes $\|A - M\|_F{}^2 = \sum_{i=1}^{m} \sum_{j}^{n} \left(A_{ij} - M_{ij}\right)^2$



Algebraic topic model: Latent Semantic Analysis (LSA)



holds orthonormal Eigenvectors of AA^T (term-term-similarity)

positive roots of the Eigenvalues (singular values) of $A^T A$

holds orthonormal Eigenvectors of $A^T A$ (doc-doc similarity)

Mapping existing docs $\vec{d}_i, \vec{d}_j$ into latent space:

$$\vec{d}_i \longmapsto D_k^{-1} {U_k}^T \vec{d}_i = \vec{d}_i' = \left(\left(D_k {V_k}^T \right)_{(i)} \right)$$

$$\vec{d}_j \longmapsto D_k^{-1} {U_k}^T \vec{d}_j = \vec{d}_j' = \left(\left(D_k {V_k}^T \right)_{(j)} \right)$$

$$\vec{q} \longmapsto D_k^{-1} {U_k}^T \vec{q} = \vec{q}' \text{ (mapping of query)}$$

… and measure cosine-similarity

Mapping of new doc d' into latent space:

$$\vec{d} \longmapsto D_k^{-1} {U_k}^T \vec{d} = \vec{d}'$$

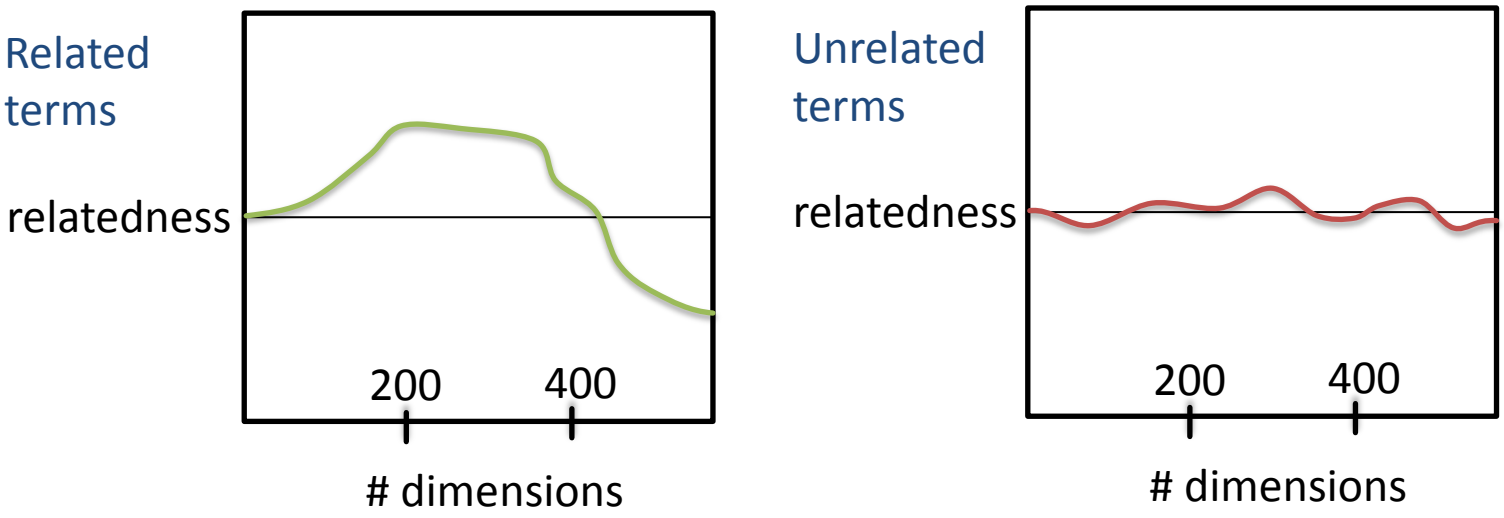
Add $\vec{d}'$ as last column of ${V_k}^T$

Impact of dimensions on term-term relatedness in LSA

- LSA: Projection of data onto top-k Eigenvectors $A \approx U_k D_k {V_k}^T$
- What about term-term similarity???

- **Typical behavior of term-term relatedness as a function of k**

  ${U_k}{U_k}^T$: $m \times m$ term-term similarity matrix



Pros and cons of LSA

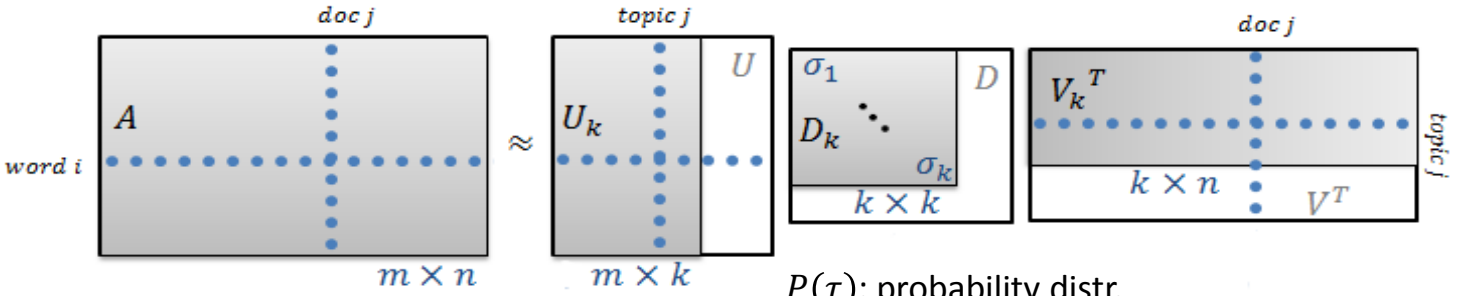
- Principled versatile mathematical model that may reveal
  - Latent topics (i.e., “semantic” clusters)
  - Word-word similarities (synonymy, polysemy)
  - Relevant docs to a query
- Computational and storage overhead
- Unrealistic assumptions on the distribution of data (LSA works best for normally distributed data)
- Principal Component Analysis (PCA) is very similar but more principled for normally distributed data; it employs data centering before applying SVD (complexity of SVD: $O(n^2k^3)$)
- How to choose k appropriately?

Probabilistic topic models: Probabilistic LSA (PLSA)

- View word-doc pairs as generated from topic: $P(wd\tau) = P(\tau)P(wd|\tau)$
- Assume word and doc independence given topic: $P(wd\tau) = P(\tau)P(w|\tau)P(d|\tau)$

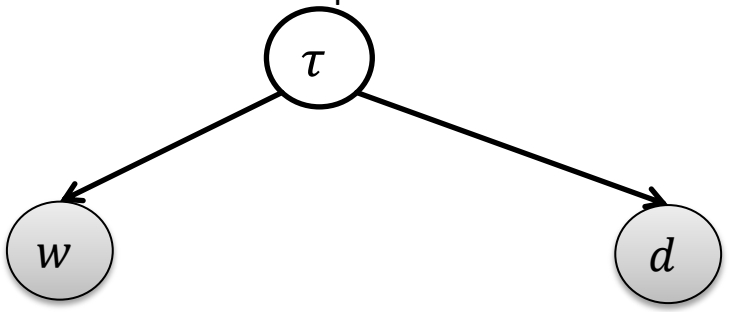
latent topic

$$P(wd) = \sum_{\tau} P(w|\tau) \; P(\tau) \; P(d|\tau)$$



$P(\tau)$: probability distr. over topics

- With multinomial $P(* \,|\tau)$

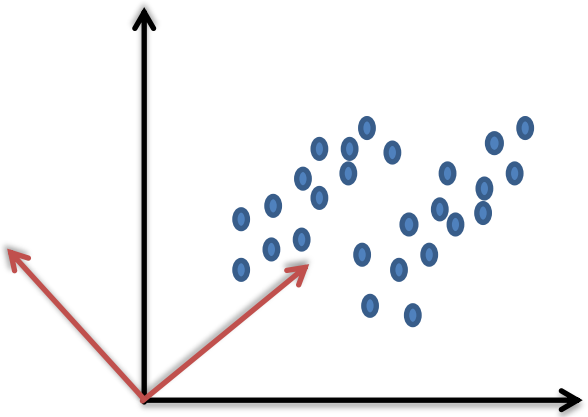


$P(w|\tau)$: probability distr. over words given topic

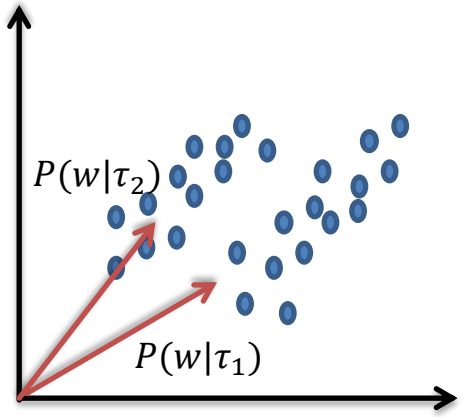
$P(d|\tau)$: probability distr. over docs given topic

Main difference between PLSA and LSA

- Matrix factorization in PLSA is **non-negative**
- Values represent probabilities (and are thus L_1-normalized)



Column vectors of U in LSA
(Eigenvectors are orthonormal)



Column vectors of U in PLSA
(non-negative matrix factorization)

Expectation Maximization for PLSA

- $P(wd) = \sum_\tau P(w|\tau)P(d|\tau)P(\tau)$
- $P\left(\{w_j, d_l\}_{1\le j\le m, 1\le l\le n}|\boldsymbol{\theta}\right) = \prod_{j,l}\sum_\tau P(w_j|\tau)P(d_l|\tau)P(\tau)$
- Maximize log of above function (i.e., log likelihood)
  - Expectation (responsibilities of latent topics under current parameters):

$$P(\tau = \tau_i|w_j, d_l) = \frac{P(\tau_i)P(w_j|\tau_i)P(d_l|\tau_i)}{\sum_\tau P(\tau)P(w_j|\tau)P(d_l|\tau)}$$

  - Maximization (update parameters based on responsibilities):

$$P(w_j|\tau_i) \propto \frac{\sum_d freq(w_j, d)P(\tau_i|w_j, d)}{\sum_{d,w} freq(w, d)P(\tau_i|w, d)},$$

$$P(d_l|\tau_i) \propto \frac{\sum_w freq(w, d_l)P(\tau_i|w, d_l)}{\sum_{d,w} freq(w, d)P(\tau_i|w, d)},$$

$$P(\tau_i) \propto \frac{\sum_d \sum_w freq(w, d)P(\tau_i|w, d)}{\sum_{d,w} freq(w, d)}$$

Tempered EM for PLSA

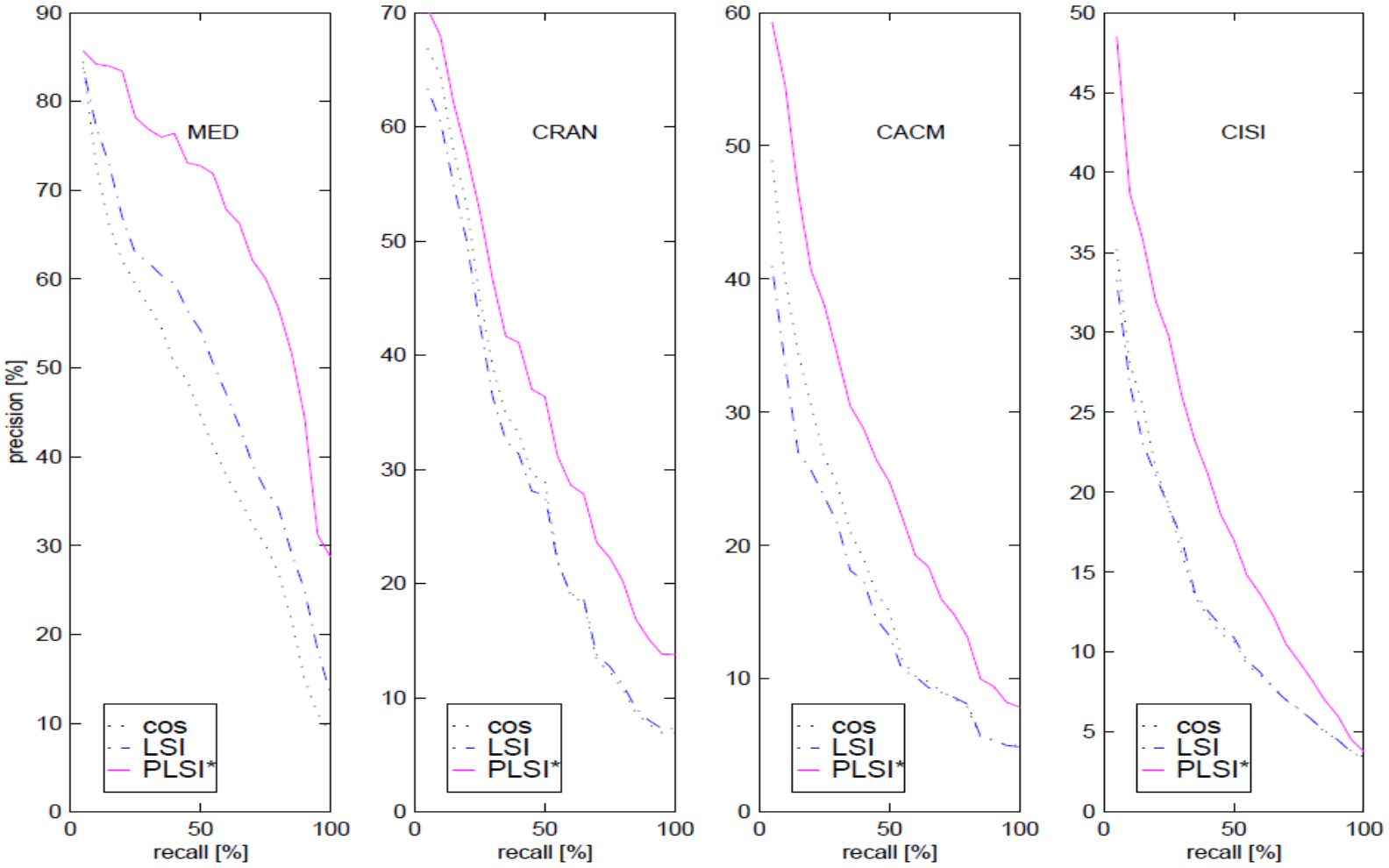
- Example of tempered EM for regularization parameter $\lambda < 1$:

$$P(\tau = \tau_i | w_j, d_l) = \frac{P(\tau_i)\left[P(w_j|\tau_i)P(d_l|\tau_i)\right]^{\lambda}}{\sum_{\tau} P(\tau)\left[P(w_j|\tau)P(d_l|\tau)\right]^{\lambda}}$$

- Dampens probabilities in M step and mitigates overfitting
- Yields faster estimation procedure compared to other methods
- How to get a good λ?
  I. Set $\lambda := 1$ and perform EM until the performance on test data deteriorates (early stopping)
  II. Decrease λ, e.g., $\lambda \leftarrow \eta\lambda, \quad \eta < 1$
  III. As long as the performance on held-out data improves continue TEM iterations at this value and stop when decreasing does not yield further improvements, otherwise go to step (II.)
  IV. Perform some final iterations using both, training and held-out data

Retrieval performance of PLSA

- Precision-recall curve for retrieval task on different corpora



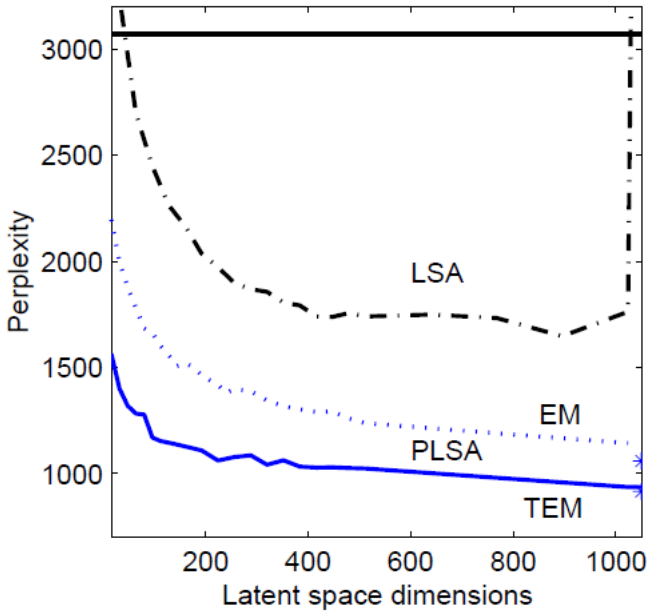
From: Thomas Hofmann, Probabilistic Latent Semantic Analysis, UAI'99

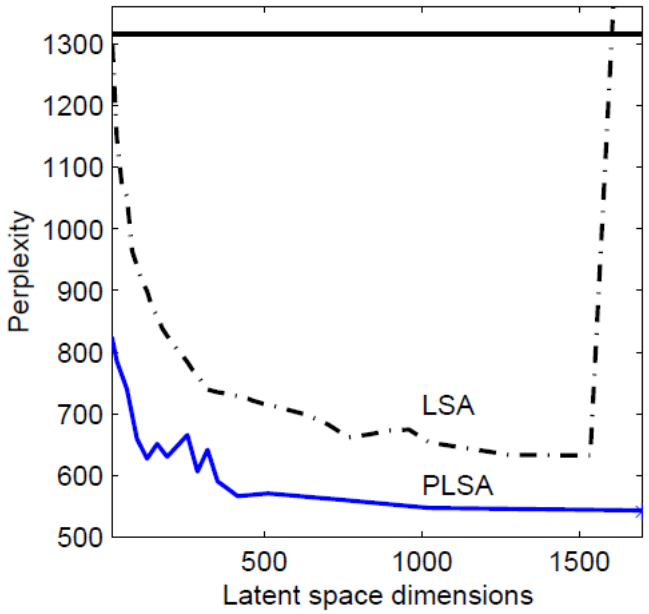
Predictive perplexity of PLSA

- **Perplexity:** Measure of generalization (i.e., how well is overfitting mitigated?)
- For different values of k, compute on a test corpus of D docs:

$$\mathrm{Perp}(\Phi) = \left(\prod_{d=1}^{D}\prod_{i=P+1}^{N_d} p(w_i \mid \Phi, w_{1:P})\right)^{\frac{-1}{\sum_{d=1}^{D}(N_d-P)}}$$

Model parameters



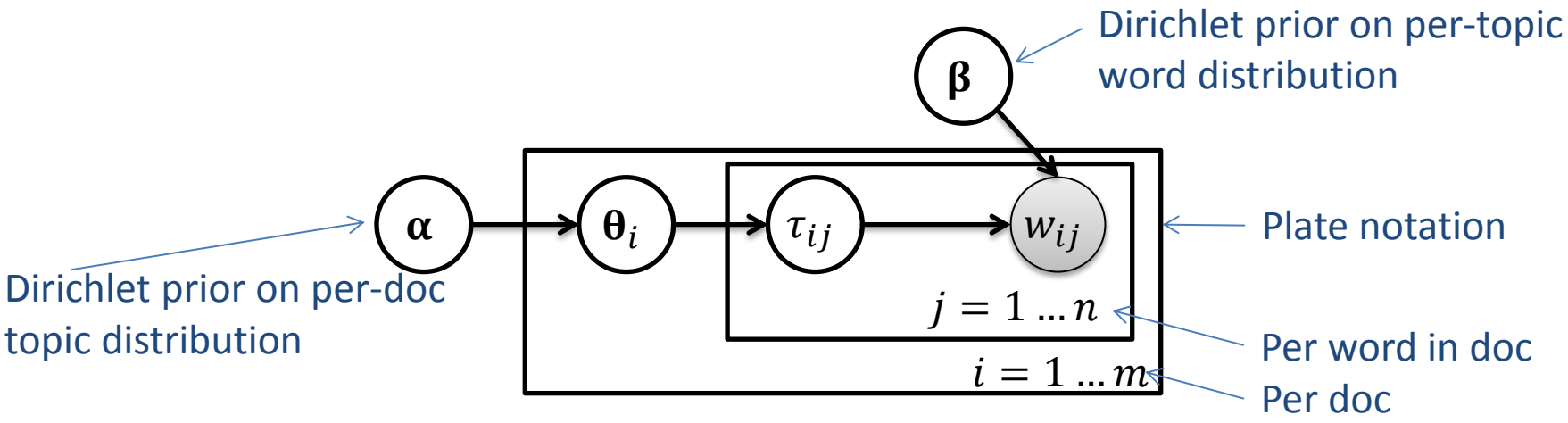


From: Thomas Hofmann, Probabilistic Latent Semantic Analysis, UAI'99

Pros and cons of PLSA

- Better empirical performance than LSA
- Mitigates overfitting better than LSA
- Different techniques can be used for estimation of probabilities (EM, tempered EM, Markov Chain Monte Carlo, …)
- Computationally expensive
- Selecting the right k still remains a problem
- Overfitting is still a problem

Probabilistic topic models: Latent Dirichlet Allocation (LDA)



- Each document is random mixture over latent topics

- **Generative model:**

  Choose $n \sim Poisson(\xi)$ // length of doc

  Choose $\boldsymbol{\theta}_i \sim Dir(\boldsymbol{\alpha})$ // where $1 \leq i \leq m$ and $\boldsymbol{\theta}_i$ is a k-dimensional vector

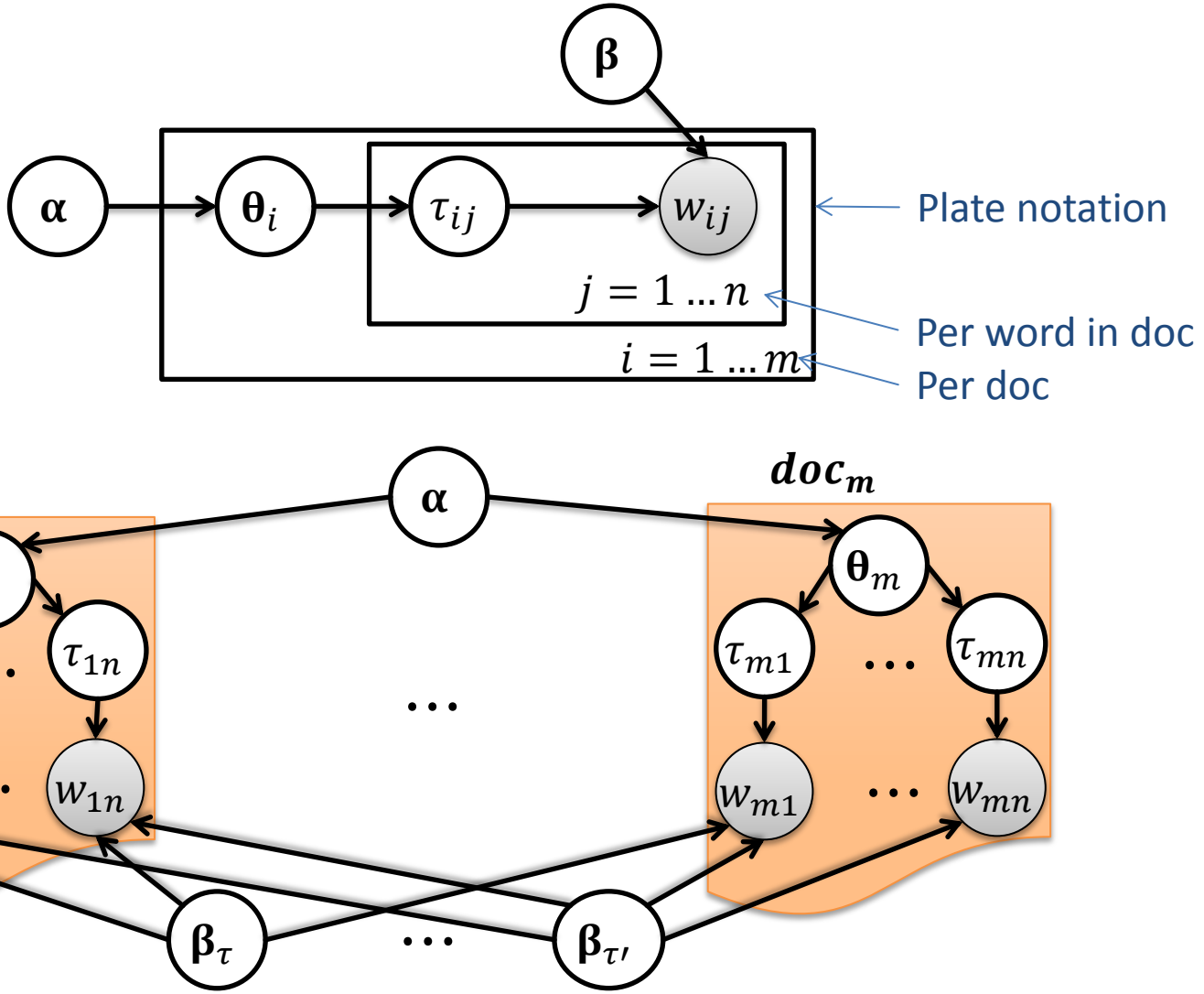
  For each w_{ij}, $j = 1 \dots n$

    Choose $\tau_{ij} \sim Multinomial(\boldsymbol{\theta}_i)$

    Choose w_{ij} from $P(w_{ij}|\tau_{ij}, \boldsymbol{\beta})$ // a multinomial distribution on τ_{ij}

    // $\boldsymbol{\beta}$ is a $k \times |W|$ matrix with $\beta_{u,v} = P(w_v = 1|\tau_u = 1)$

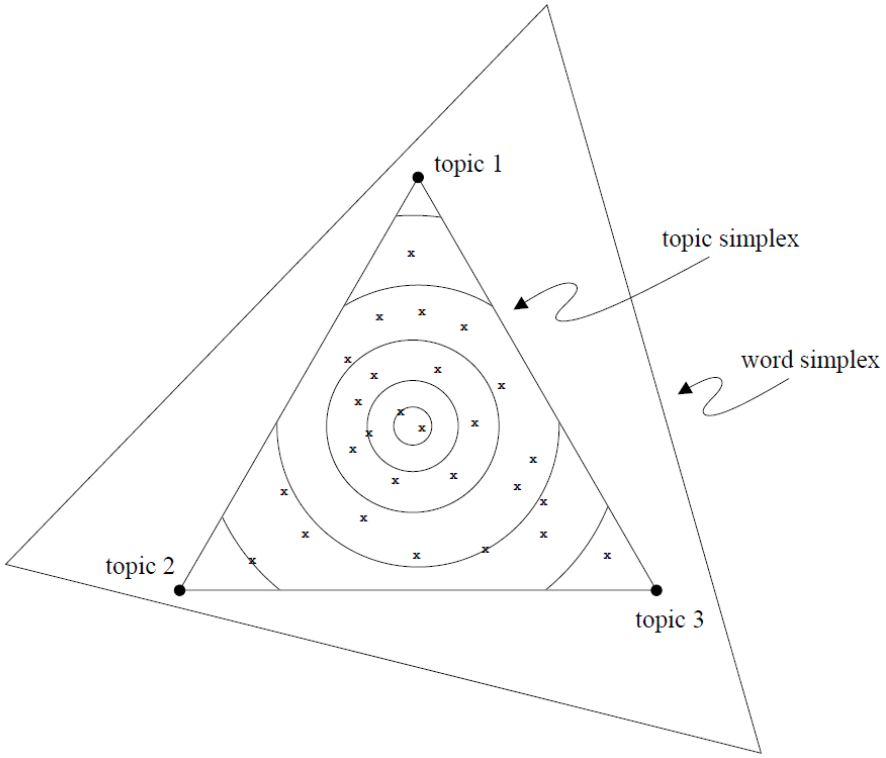
Grounded version of the LDA model



Geormetric visualization of LDA

- k topics are placed in a $(k-1)$-dimensional probability simplex (spanned by the Dirichlet distribution)

$$P(\theta_{j1}, \ldots, \theta_{jk}) = \frac{1}{Beta(\boldsymbol{a})} \prod_{i=1}^{k} {\theta_{ji}}^{a_i - 1}$$

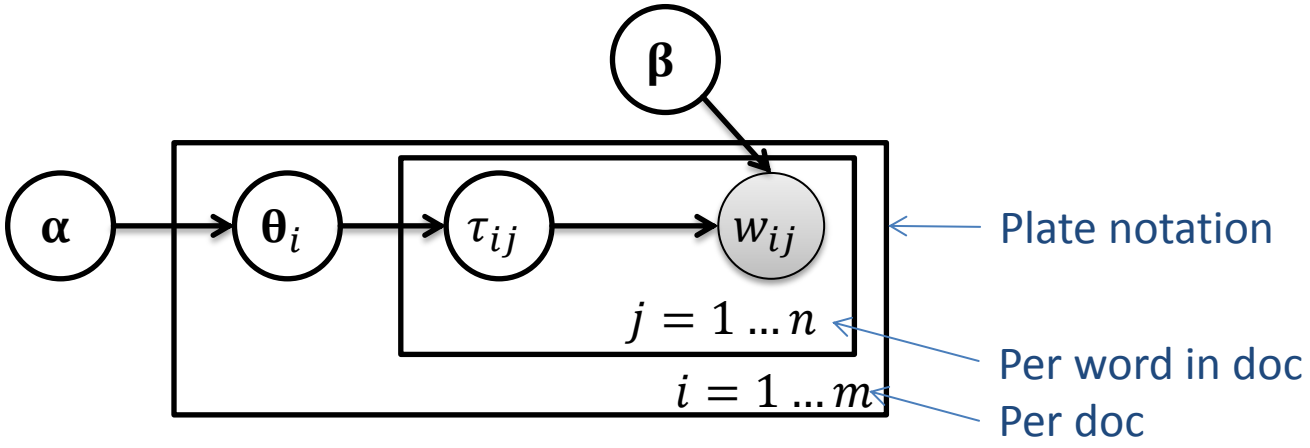


PLSA can place docs only at certain points in topic simplex

LDA imposes smooth distribution on topic simplex and can place docs at arbitrary points in the simplex

Source: D. Bleiet al.: Latent Dirichlet Allocation, JMLR 2003

Estimation of LDA parameters (1)



- By Markov Condition: In a Bayesian network a node is conditionally independent of all its non-descendants given its parents

$$P(\boldsymbol{d}_1, \dots, \boldsymbol{d}_m, \boldsymbol{\tau}, \boldsymbol{\theta} | \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{m} P(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) \prod_{j=1}^{n_i} P(\tau_{ij} | \boldsymbol{\theta}_i) P(w_{ij} | \tau_{ij}, \boldsymbol{\beta})$$

Observed word vectors for each doc

- Target function (analytically intractable)

$$P(\boldsymbol{d}_1, \dots, \boldsymbol{d}_m | \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{m} \int P(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) \left(\prod_{j=1}^{n_i} \sum_{\tau_{ij}} P(\tau_{ij} | \boldsymbol{\theta}_i) P(w_{ij} | \tau_{ij}, \boldsymbol{\beta}) \right) d\boldsymbol{\theta}_i$$

Estimation of LDA parameters (2)

- **Inference problem**

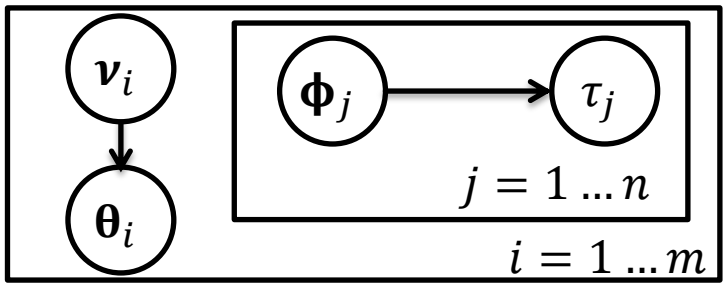
  Given doc $\boldsymbol{d} = \{w_1, \ldots, w_n\}$, what topics does it belong to?

  Estimate hidden parameters:

$$P(\boldsymbol{\theta_d}, \boldsymbol{\tau_d} | \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{P(\boldsymbol{\theta_d}, \boldsymbol{\tau_d}, \boldsymbol{d} | \boldsymbol{\alpha}, \boldsymbol{\beta})}{P(\boldsymbol{d} | \boldsymbol{\alpha}, \boldsymbol{\beta})}$$

- **Variational inference:** Approximate $P(\boldsymbol{\theta_d}, \boldsymbol{\tau_d} | \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ with simpler distribution $Q(\boldsymbol{\theta_d}, \boldsymbol{\tau_d} | \boldsymbol{\nu_d}, \boldsymbol{\phi})$ and find:

$$(\boldsymbol{\nu'_d}, \boldsymbol{\phi'_d}) = \underset{\boldsymbol{\nu_d}, \boldsymbol{\phi_d}}{\operatorname{argmin}} KL(Q(\boldsymbol{\theta_d}, \boldsymbol{\tau_d} | \boldsymbol{\nu_d}, \boldsymbol{\phi_d}) \parallel P(\boldsymbol{\theta_d}, \boldsymbol{\tau_d} | \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\beta}))$$



Dirichlet Multinomial

Optimization problem that yields tight lower-bound on log likelihood!

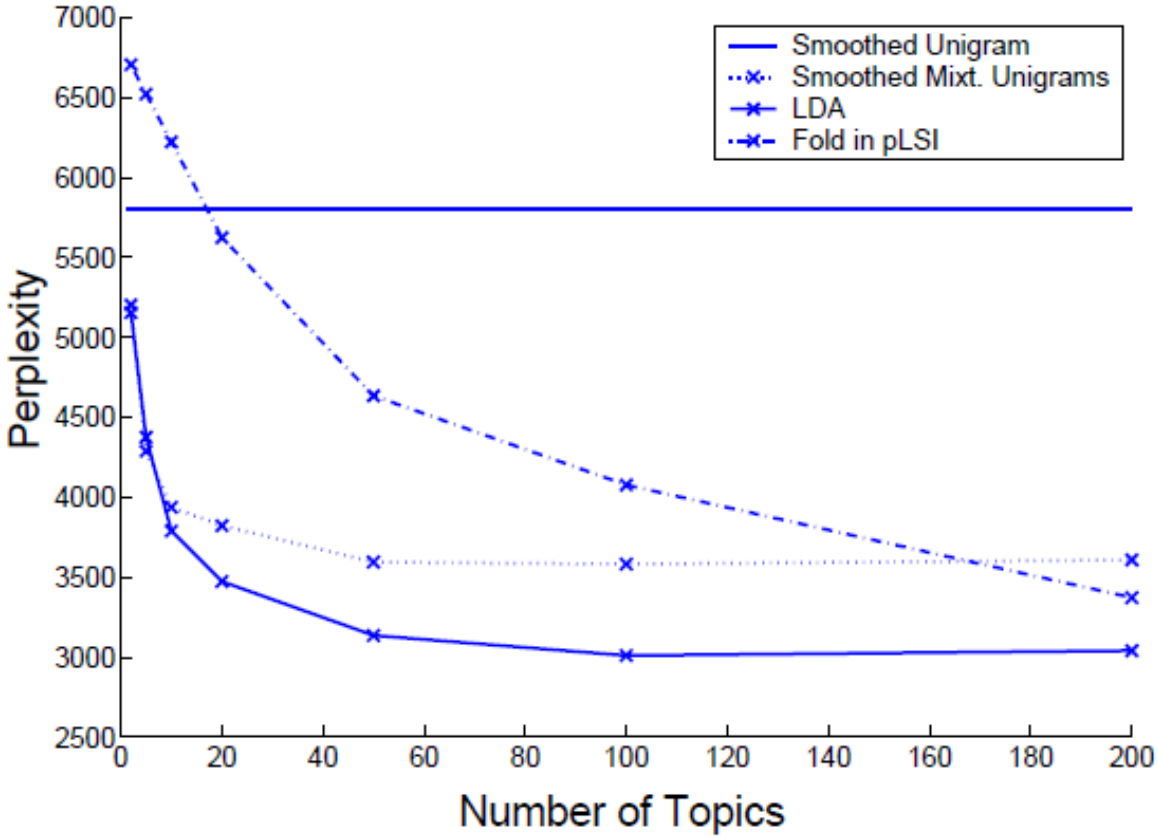
LDA example

NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

From: D. Blei et al.: Latent Dirichlet Allocation, JMLR 2003

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Perplexity of LDA



Pros and cons of LDA

- Principled probabilistic model and inference (widely used, e.g., in Information Retrieval, Machine Learning, Natural Language Processing, …)
- Better empirical behaviour than PLSA
- Mitigates overfitting better than other topic models
- High computational complexity, even with approximate inference

Summary

- Latent topic models

LSA

- Deerwester, S., et al., **Improving Information Retrieval with Latent Semantic Indexing**, Proceedings of the 51st Annual Meeting of the American Society for Information Science 25, 1988, pp. 36–40.

PLSA

- T. Hofmann, **Probabilistic Latent Semantic Analysis**, Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99), 1999.

LDA

- D. Blei, A. Ng, and M. Jordan, **Latent Dirichlet Allocation**, Journal of Machine Learning Research, 3:993-1022, January 2003.

Also see Wikipedia articles on LSA, PLSA and LDA