LSA, PLSI, and LDA

A Brief Introduction

Representation

- RG (Niu Li-Qiang)
  - LSA → word2vec → GloVec

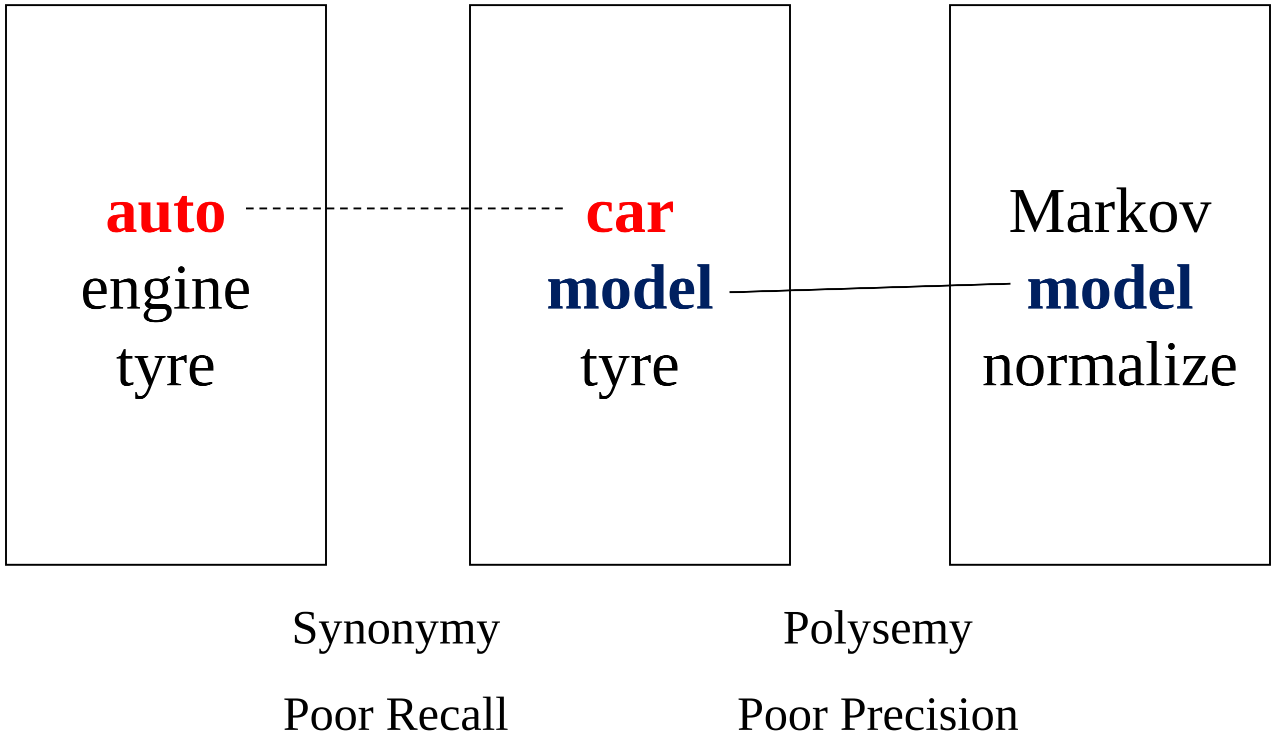
- 本次
  - LSA→PLSA→LDA

Timeline & Outline

- 1990， LSA
  - *Indexing by latent semantic analysis*. Deerwester, Dumais et al. JASIS
- 1999， PLSI
  - *Probabilistic Latent Semantic Indexing*. Hofmann. SIGIR
- 2003， LDA
  - *Latent Dirichlet Allocation*. Blei, Ng, Jordan. JMLR

Two Problems in Natural Language

- Synonymy
  - 一义多词
- Polysemy
  - 一词多义



The Setting

- Corpus
  - set of documents
  - D = {d_1, …, d_N}
- Vocabulary
  - set of words
  - W = {w_1, …, w_M}
- Term-Doc Matrix
  - occurrence of words in docs
  - A_ij = n(w_i, d_j)

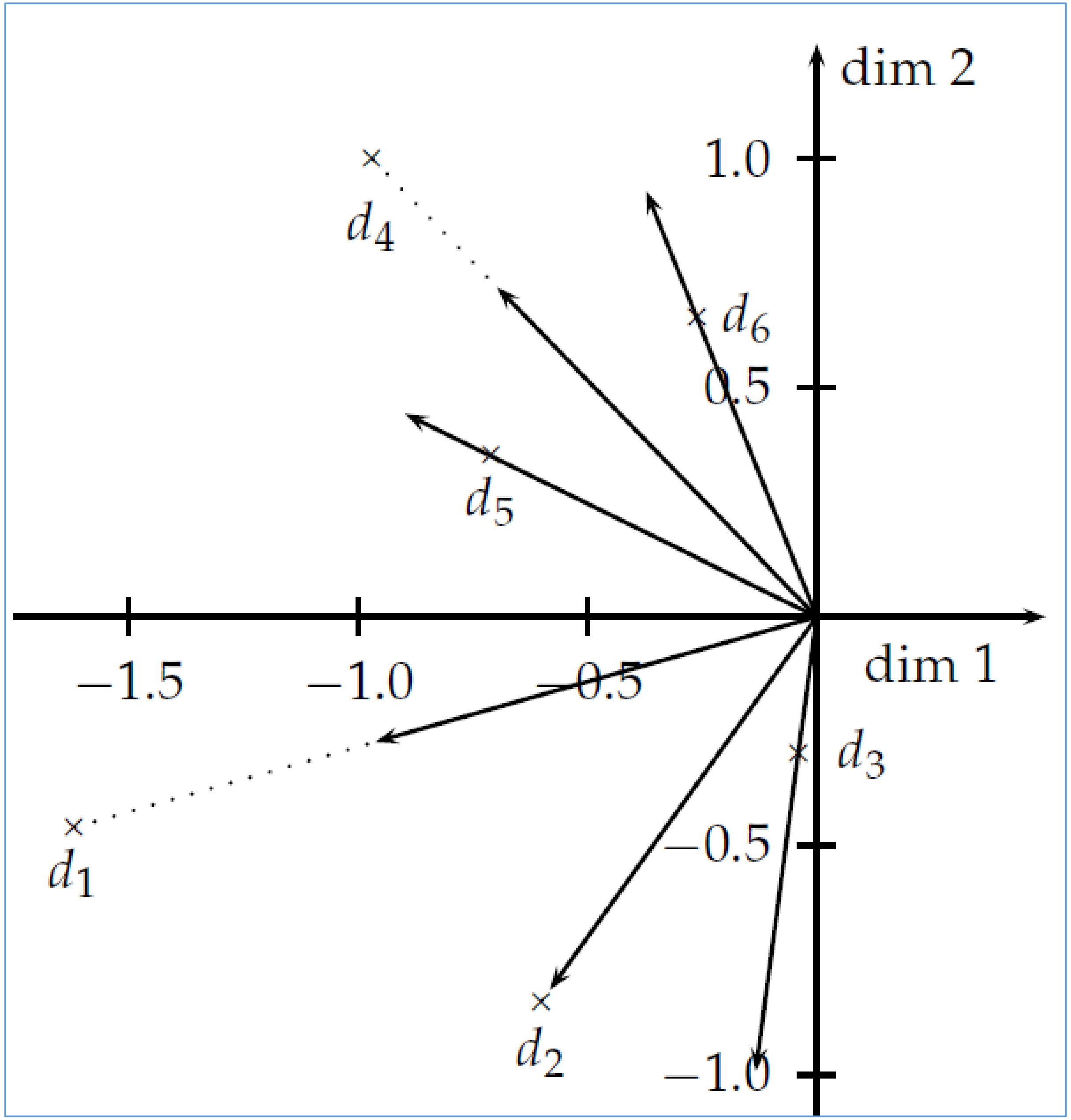
Term-Doc	D_1	D_2	D_3
auto	1		
engine	1		
tyre	1	1	
car		1	
model		1	1
markov			1
normalize			1

<u>NOTE: Each example has its own docs and terms which are from different references.</u>

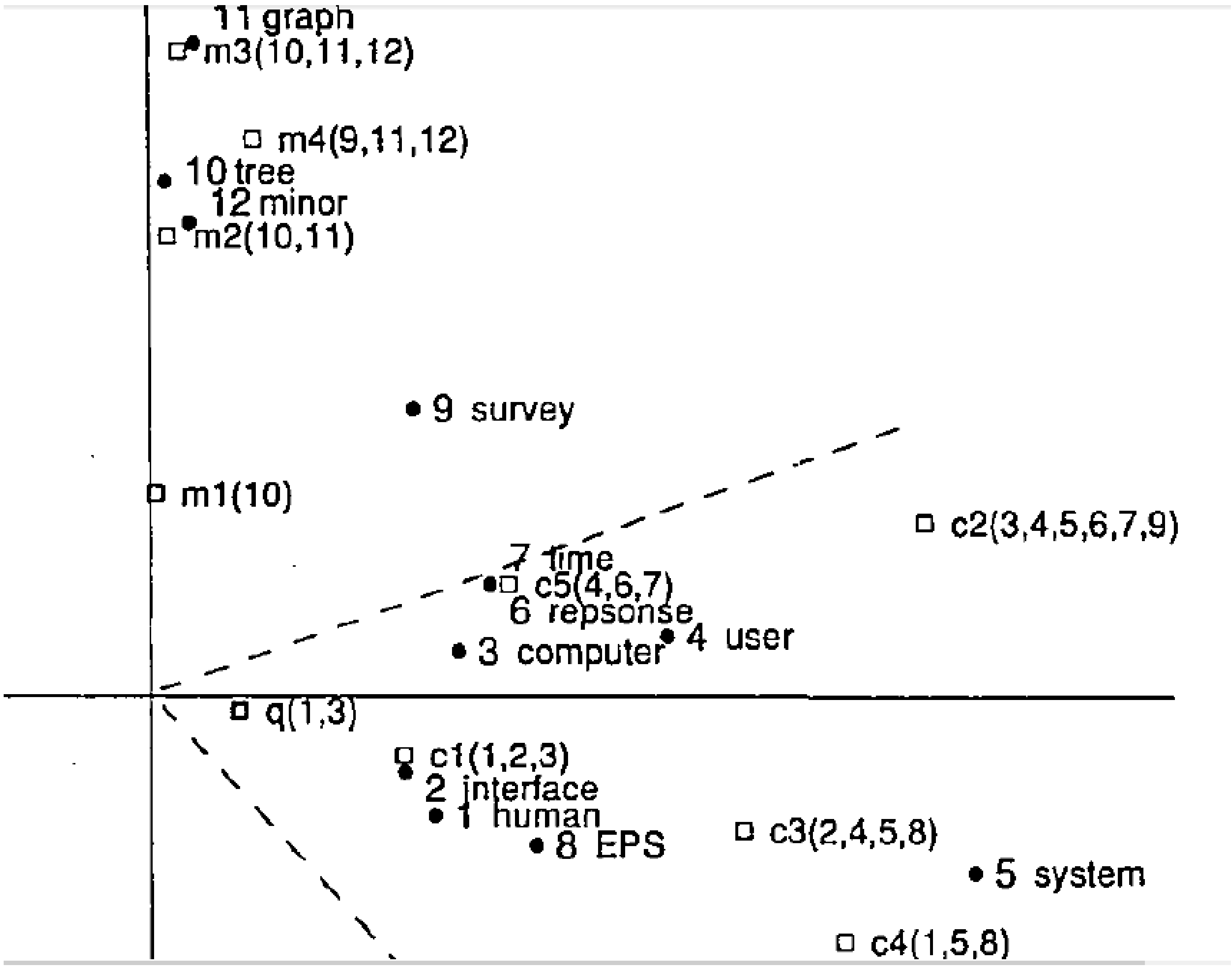
I

LSA: Key ideas

- Mapping terms and docs into a **latent semantic space**



1	Human
2	Interface
3	Computer
4	User
5	System
6	Response
7	Time
8	EPS
9	Survey
10	Tree
11	Graph
12	Minor



LSA: Technical Details (1)

- Matrix Diagonalization Theorem (Eigen Decomposition)

$$S^{-1}AS = \Lambda = diag(\lambda_1, \ldots, \lambda_n)$$
$$A \in R^{n \times n}, S : Eigenvectors, \Lambda : Eigenvalues$$

- Symmetric Diagonalization Theorem (Spectral Theorem)

$$A = Q\Lambda Q^{-1} = Q\Lambda Q^T$$
$$A \in S^{n \times n}, Q : Orthogonal$$

- Singular Value Decomposition (SVD)

$$C = U\Sigma V^T = \sum_i \sigma_i u_i v_i^T, \ \Sigma : (C)\ Singular Values$$
$$U : \left(CC^T\right) Eigenvectors, V : \left(C^T C\right) Eigenvectors$$

LSA: Technical Details (2)

- Eckart – Young Theorem (1936)

$$\min_{Z:r(Z)=k} \|C - Z\|_F^2 = \|C - C_k\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2$$

- Keep top k singular values(Optimal in the sense of L2-norm)
- Term-Term correlation: C C^T (C: term-doc matrix)
- Doc-Doc correlation: C^T C
- Query representation: $V_q = C_{\mathrm{q}}^T U \Sigma^{-1}, V_q \in R^{1\times K}, C_q \in R^{1\times m}$

Aside: PCA and SVD

- The principal components of matrix X are rows of orthogonal matrix P such that the covariance C_Y of Y≡PX is diagonal
  - Rows are pre-centered $C_Y = \frac{1}{n} Y Y^T = P C_X P^T$
  - Spectral Theorem provides

$$C_Y = P C_X P^T = P(Q \Lambda Q^T)P^T = Q^T (Q \Lambda Q^T) Q = \Lambda, P = Q^T$$

  - Principal components are the eigenvectors of covariance
- Finding PCA via SVD
  - Construct Y $Y \equiv \frac{1}{\sqrt{n}} X^T, Y^T Y = C_X$
  - Perform SVD $Y = U \Sigma V^T, Y^T Y = V \Sigma^2 V$
  - Q = V

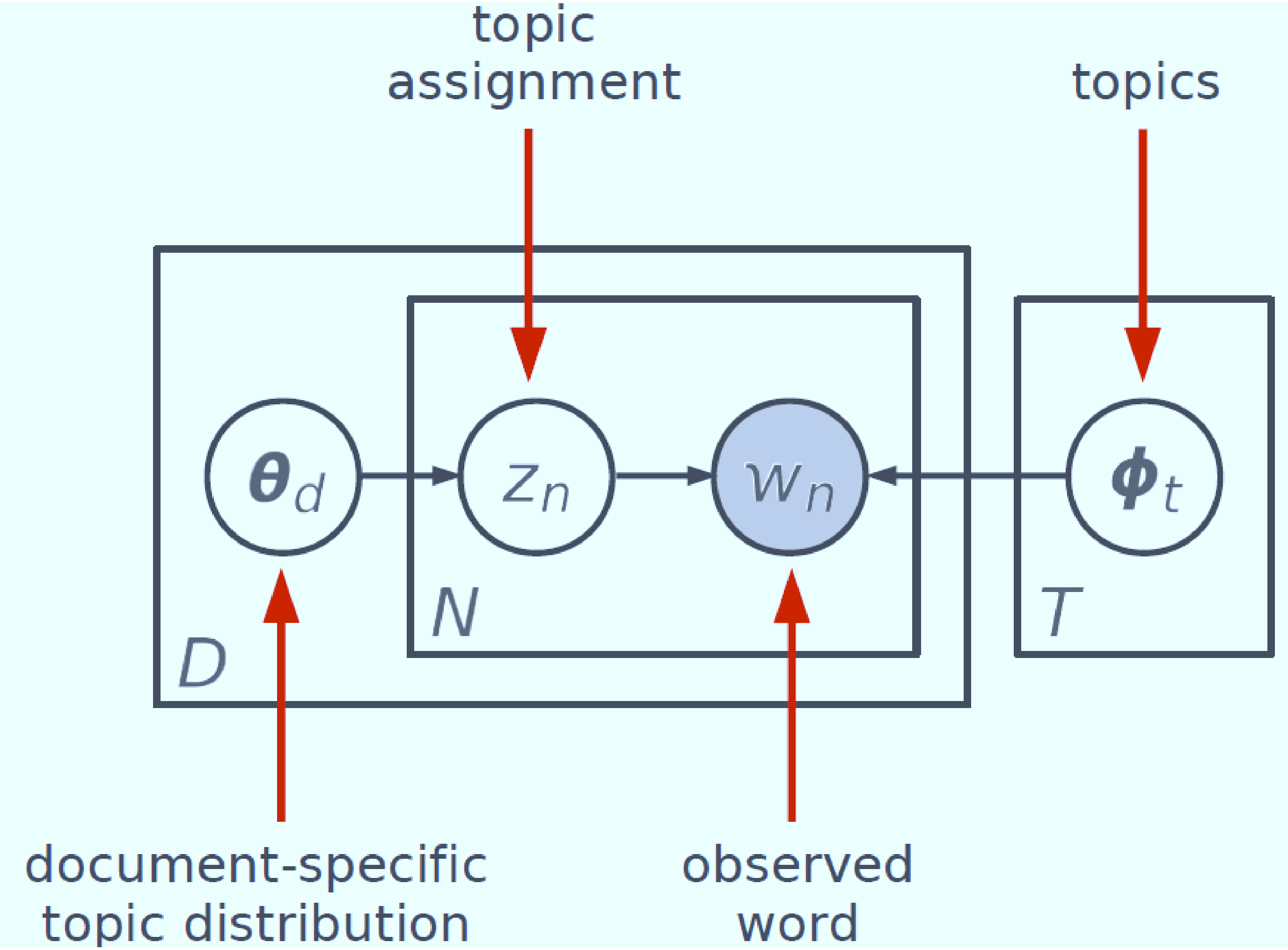
From LSA to PLSI

- LSA assumption on the data : normally distributed
  - Optimize Sum-Squares-Error ([**Eckart-Young 1936**]())
- LSA have negative entries
  - Orthogonal, not Non-negative
- *Count data (e.g. Text)*
  - Normal distribution is not appropriated; maybe multinomial better
- Latent semantic space
  - No probabilistic interpretation
- <u>Linear algebra to Probabilistic modeling</u>

II

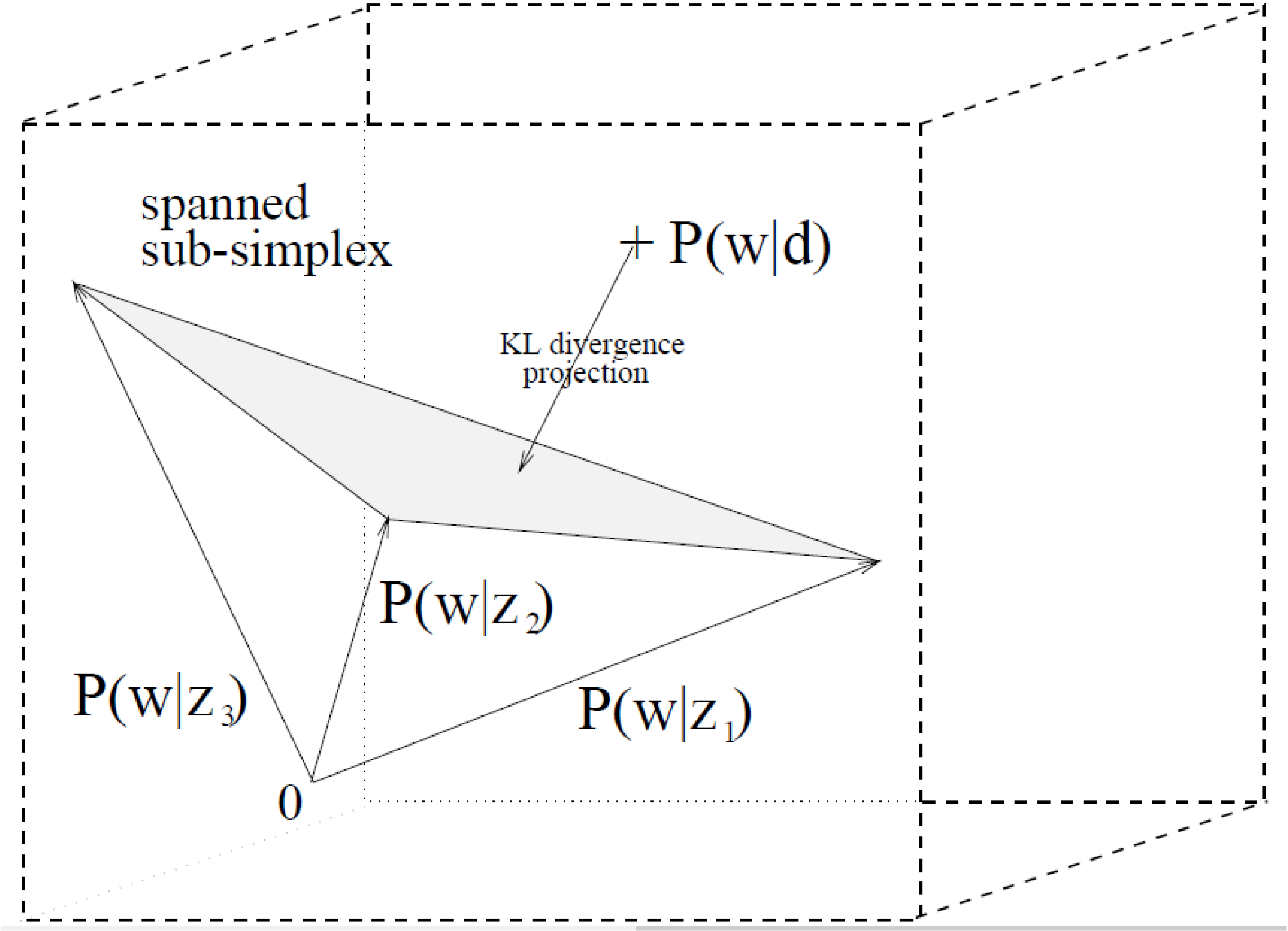
PLSI: Key Ideas

- Expressing words and documents in terms of probabilistic topics
    - Z: latent class/aspect/topic
- Generative probabilistic models of text corpora



$$P(w \mid d) = \sum_{z=1}^{K} P(z \mid d) P(w \mid z)$$

- *P(w|d)* are approximated by a *multinomial* representable as a convex combination of the class-conditionals *P(w|z)*



PLSI: An Example

- Polysemy: matrix
  - Linear algebra: 矩阵
  - Biology cell: 基质
- *LSA (SVD) helplessness*

- Making topics for central bridge
  - Doc-topic
  - Topic-term

"matrix 1"	"matrix 2"
robust	manufactur
MATRIX	cell
eigenvalu	part
uncertainti	MATRIX
plane	cellular
linear	famili
condition	design
perturb	machinepart
root	format
suffici	group

PLSI: Technical Details

- Joint Probability Model

$$L = \sum_d \sum_w n(d,w) \log P(\mathrm{d},\mathrm{w}) \qquad P(\mathrm{d},\mathrm{w}) = \sum_z P(z)P(d \mid z)P(w \mid z)$$

- E-step

$$P(\mathrm{z} \mid \mathrm{d},\mathrm{w}) = P(z)P(d \mid z)P(w \mid z) / Z \quad Z = \sum_{z'} P(z')P(d \mid z')P(w \mid z')$$

- M-step

$$P(\mathrm{z}) = \sum_{d,w} n(d,w)P(z \mid d,w) / N, \; N = \sum_{d,w} n(d,w)$$

$$\mathrm{P}(\mathrm{w} \mid \mathrm{z}) = \sum_d n(d,w)P(z \mid d,w) / W, \; \mathrm{W} = \sum_{w',d} n(d,w')P(z \mid d,w')$$

$$\mathrm{P}(d \mid \mathrm{z}) = \sum_w n(d,w)P(z \mid d,w) / D, \; D = \sum_{d',w} n(d',w)P(z \mid d',w)$$

Comparing PLSA with LSA

- LSA vs. PLSA
  - U: P(w|z)
  - V: P(d|z)
  - Sigma: P(z)

$$C = U \Sigma V^T = \sum_i \sigma_i u_i v_i^T$$

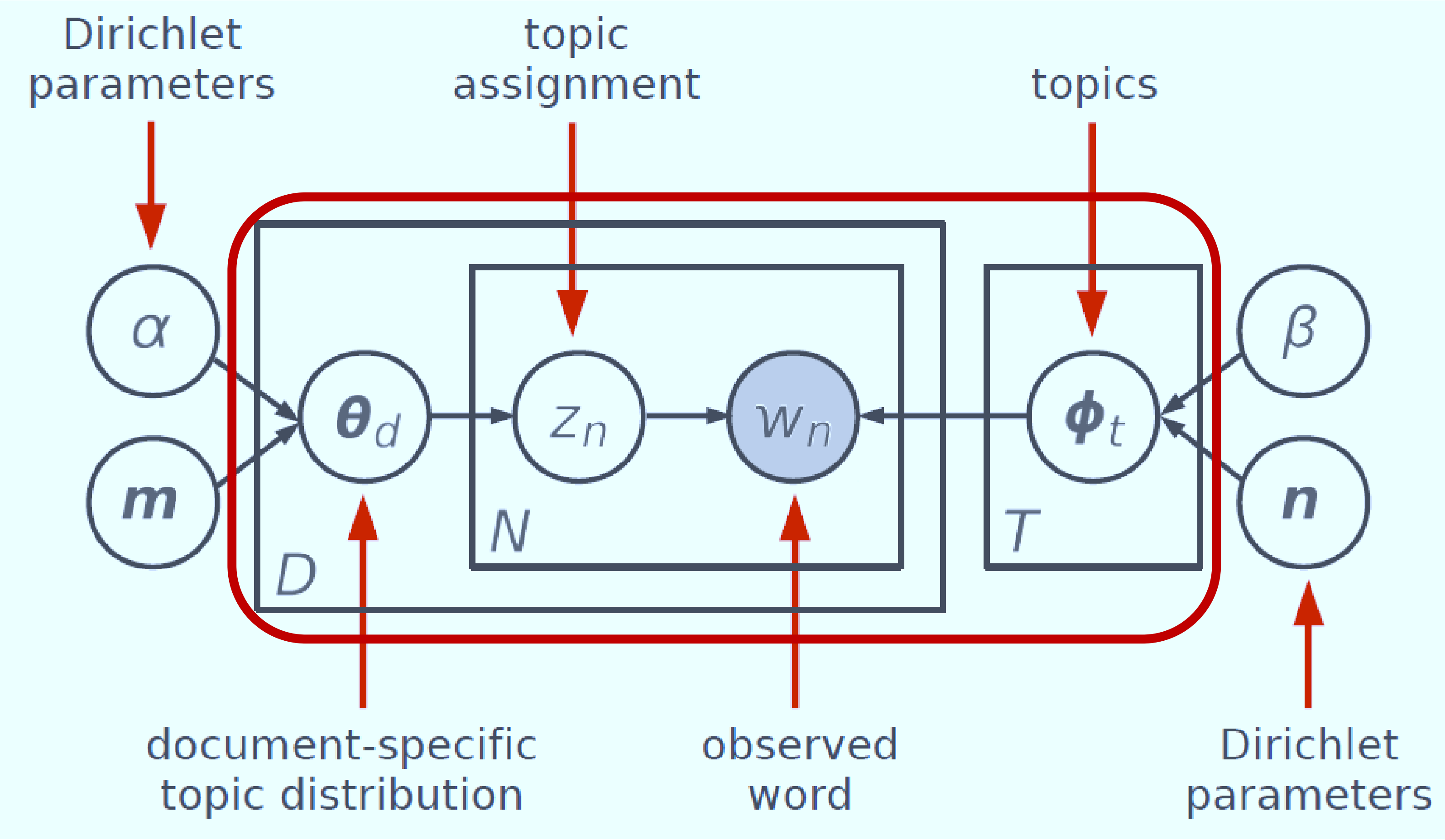
$$P(\mathrm{d}, \mathrm{w}) = \sum_z P(z) P(w \mid z) P(d \mid z)$$

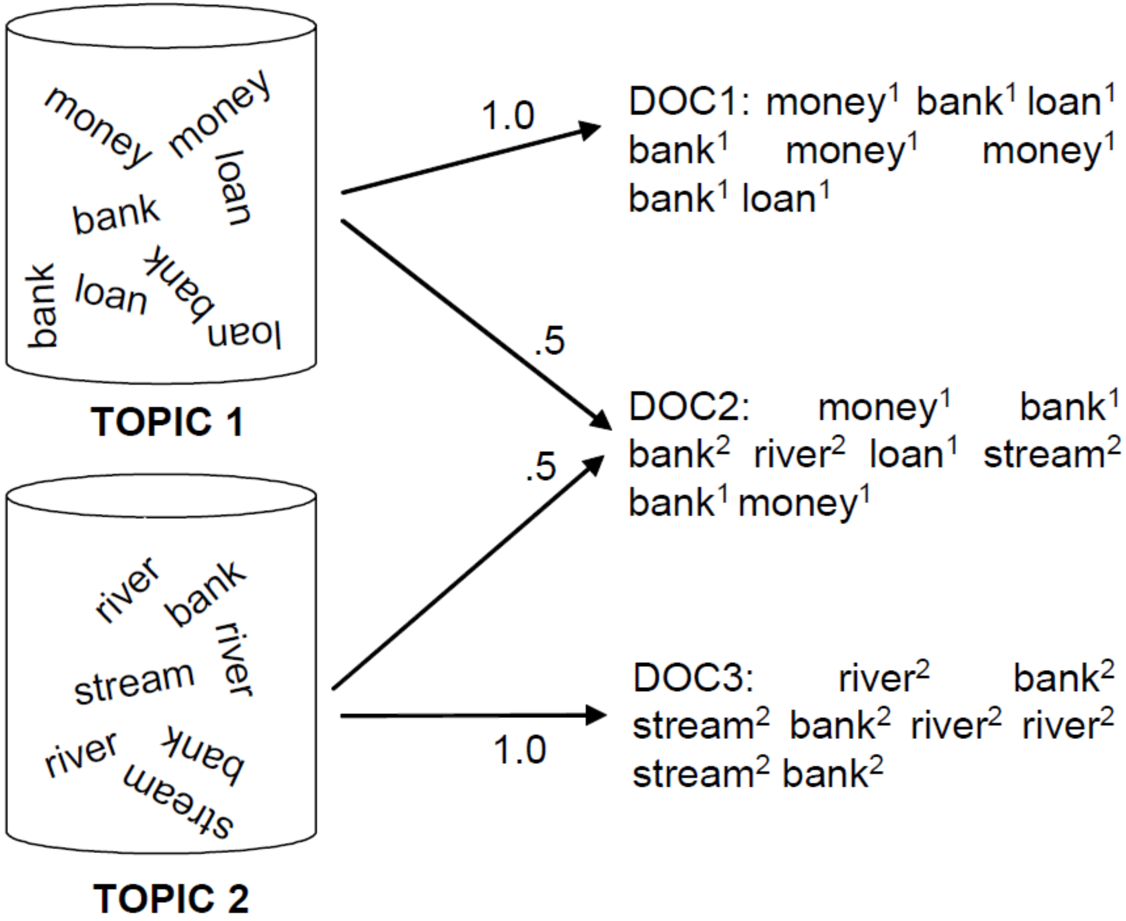
- Linear algebra (SVD) vs. Probabilistic modeling (EM)

III

LDA: Key Ideas

- Generative probabilistic models of text corpora
  - Same as PLSA
- Three-level Pr. Model
  1. Corpus: **alpha**, beta
  2. Doc: theta
  3. Term: z, w





Mixture components

Mixture weights

Bayesian approach: use priors

Mixture weights ~ Dirichlet(α)

Mixture components ~ Dirichlet(β)

LDA: Technical Details (1)

- The complete probability model

$$p(W \mid \Theta, \Phi) = \prod_{m=1}^{M} p(\vec{w}_m \mid \vec{\theta}_m, \Phi) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} p(w_{m,n} \mid \vec{\theta}_m, \Phi)$$

$$p(w_{m,n} \mid \vec{\theta}_m, \Phi) = \sum_{k=1}^{K} p(w_{m,n} \mid \vec{\varphi}_k) \, p(z_{m,n} = k \mid \vec{\theta}_m)$$

- Gibbs Sampling

$$p(z_i = k \mid \vec{z}_{-i}, \vec{w}) \propto \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,-i}^{(t)} + \beta_t} (n_{m,-i}^{(k)} + \alpha_k)$$

$$n_{u,-i}^{(v)} = n_u^{(v)} - \delta(u - u_i)$$

$n_k^{(t)}$: number of times that term t has been observed with topic k

$n_m^{(k)}$: number of times that topic k has been observed with doc m

LDA: Technical Details (2)

- while not finished do
  - for all documents *m* in [1, M] do
    - for all words *n* in [1, Nm] in document *m* do
      - decrement counts and sums:

        $$n_m^{(k)} -= 1, n_m -= 1; n_k^{(t)} -= 1, n_k -= 1$$

      - sample topic index:

        $$k' \sim p(z_i \mid \vec{z}_{-i}, \vec{w})$$

      - increment counts and sums:

        $$n_m^{(k')} += 1, n_m += 1; n_{k'}^{(t)} += 1, n_{k'} += 1$$

LDA: Technical Details (3)

- Multinomial parameters

$$\theta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k}$$

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_t}$$

Three-level models

- Unigram model

$$P(w) = \prod_n p(w_n)$$

- Mixture of unigrams

$$P(w) = \sum_z p(z) \prod_n p(w_n \mid z)$$

- PLSA

$$P(d, w_n) = P(d) \sum_z p(d \mid z) p(w_n \mid z)$$

- LDA

$$p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_n \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta$$