

Topic Modeling



Zengchang Qin (PhD)
qinzengchang@keep.com

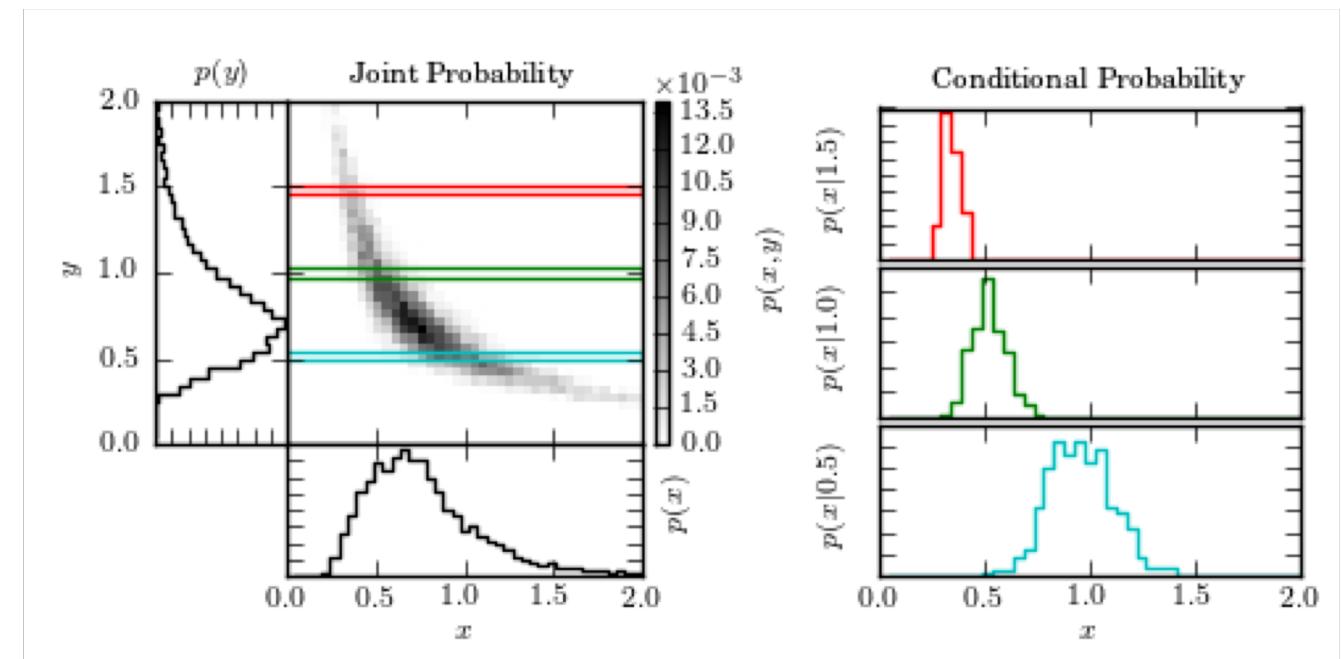
Generative Learning Models

Discriminative and Generative Learning

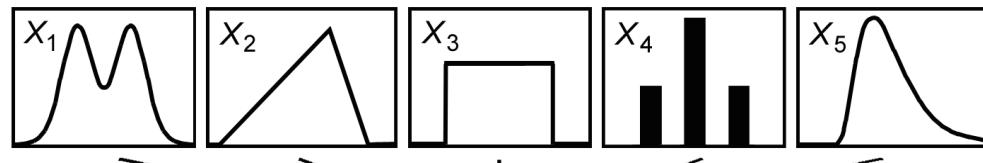
1. Algorithms that try to learn $p(x|y)$ directly from the space of inputs X to the labels {0,1}, are called discriminative learning algorithms. Aim to minimize the cost function.

2. For generative learning algorithms, we usually learn $p(x, y)$, if y indicates whether an example is a dog (0) or an elephant (1), then $p(x|y = 0)$ models the distribution of dogs' features. Aim to recreate how the observed data are generated following some hidden patterns (or rules).

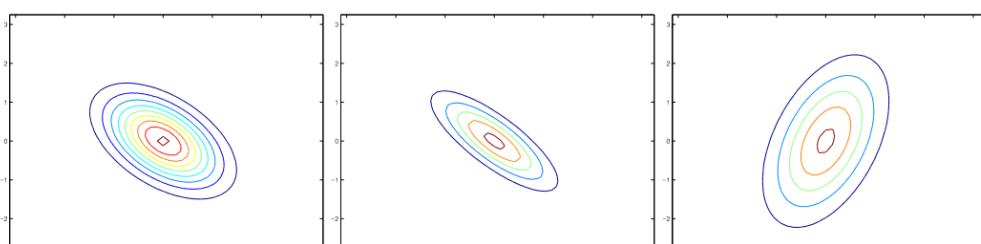
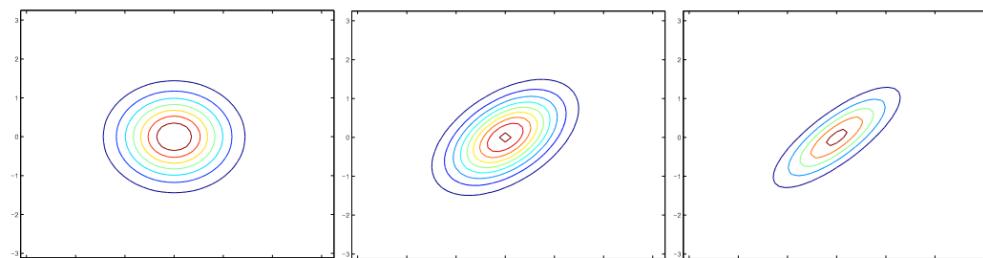
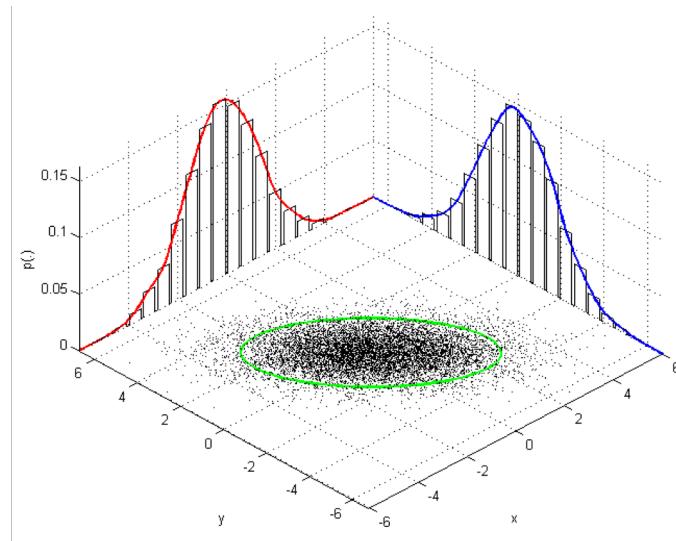
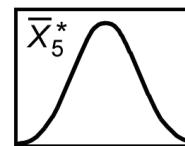
3. Based on my generation assumptions, which category is most likely to generate this signal? A discriminative algorithm does not care about how the data was generated.



Gaussian (Central Limit Theorem)



$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}; \quad .\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

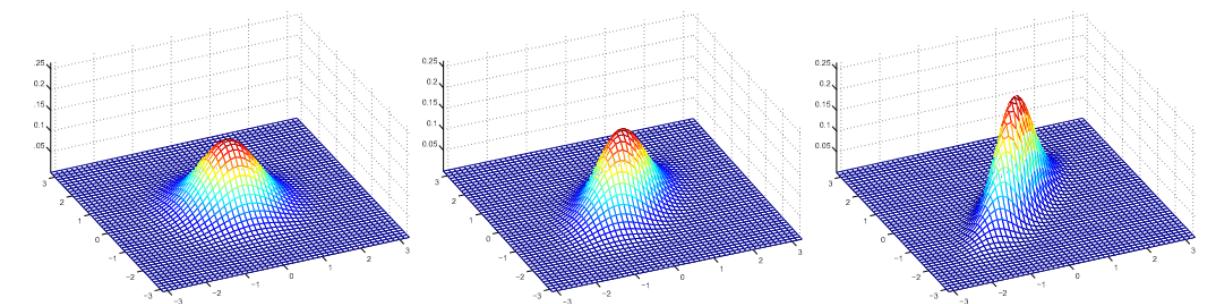
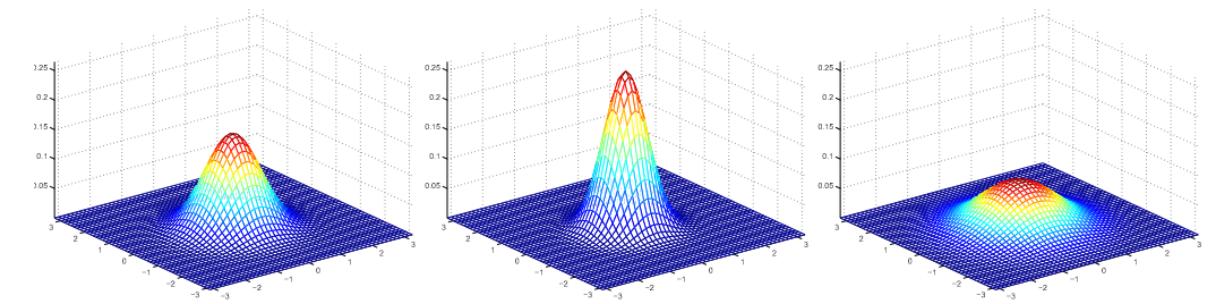
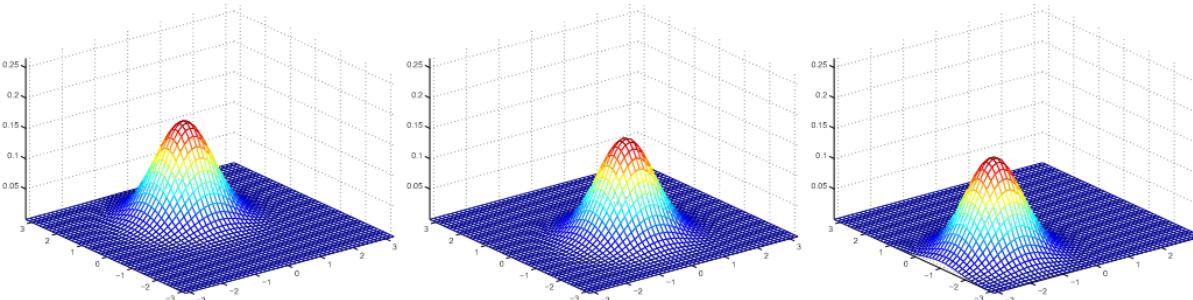


$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}; \quad .\Sigma = \begin{bmatrix} 3 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

Multivariate Gaussian

The multivariate Gaussian distribution, is parameterized by a mean vector $\mu \in R^n$ and a covariance matrix $\Sigma \in R^{n \times n}$ where $\Sigma \geq 0$ is symmetric and positive semi-definite. $|\Sigma|$ denotes the determinant of Σ :

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

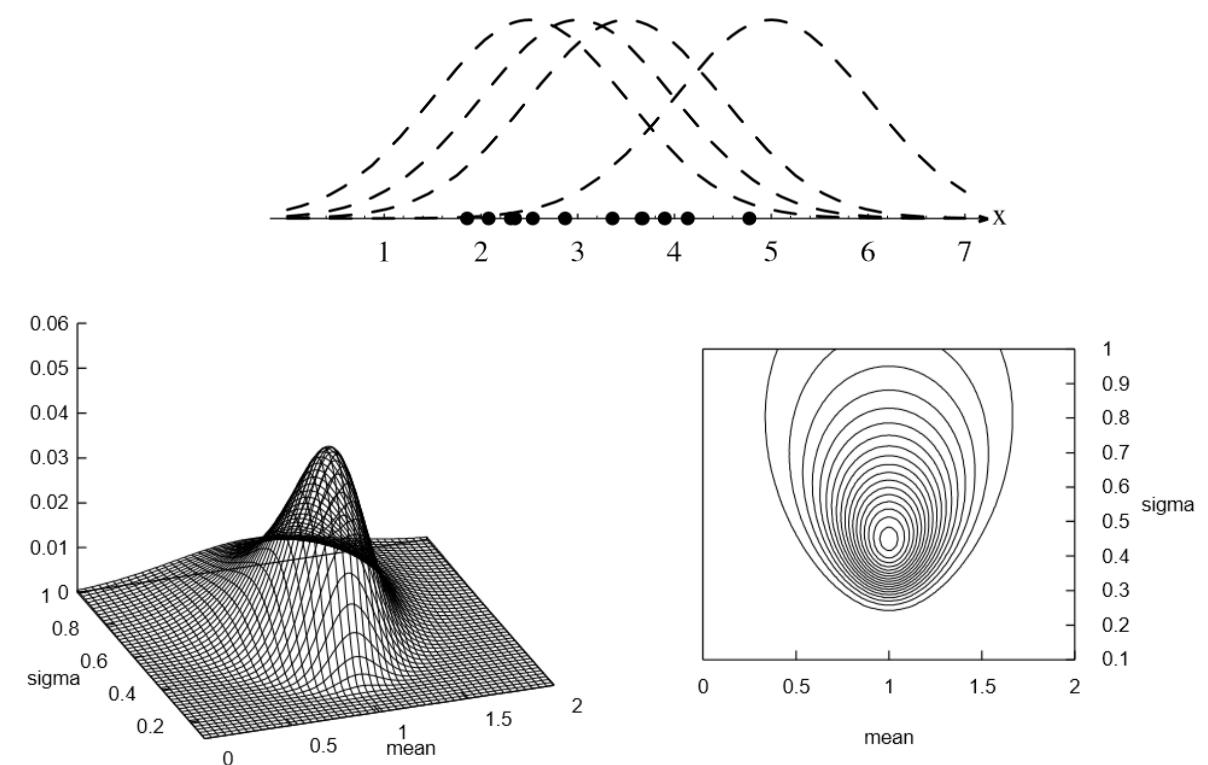
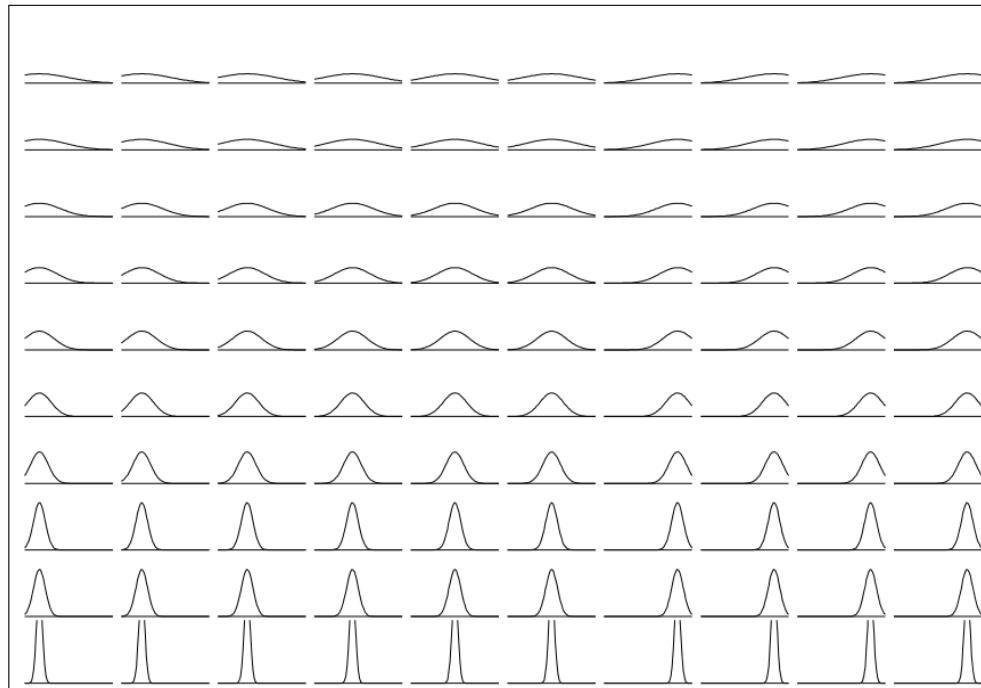


$$\mu = \begin{bmatrix} 1 \\ 0 \end{bmatrix}; \quad \mu = \begin{bmatrix} -0.5 \\ 0 \end{bmatrix}; \quad \mu = \begin{bmatrix} -1 \\ -1.5 \end{bmatrix}$$

Maximum Likelihood

We change the mean and std in order to obtain the maximum likelihood of $\{x_1, x_2 \dots x_N\}$ (e.g. N=5):

Q5: How to calculate the likelihood?



Gaussian Discriminant Analysis

Given y is a binary random variable follows Bernoulli distribution parameterized with ϕ :

$$\begin{aligned} y &\sim \text{Bernoulli}(\phi) \\ x|y=0 &\sim \mathcal{N}(\mu_0, \Sigma) \\ x|y=1 &\sim \mathcal{N}(\mu_1, \Sigma) \end{aligned}$$

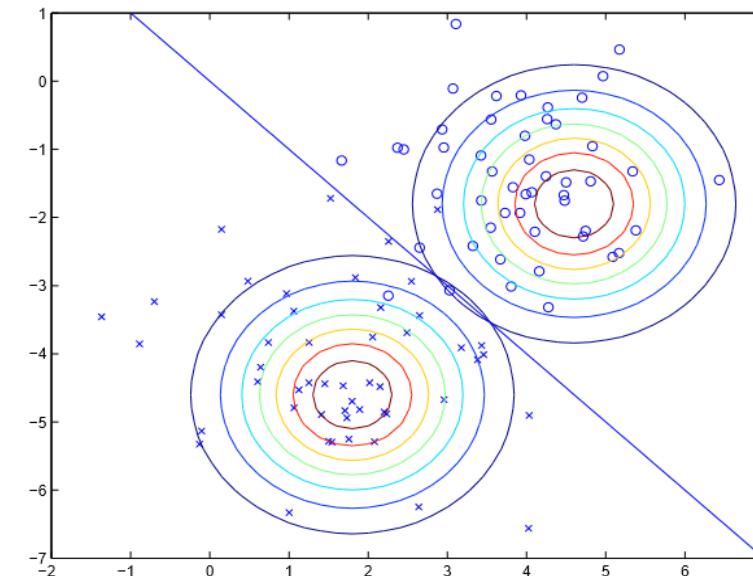
Specifically:

Q5: Generative or discriminative ?

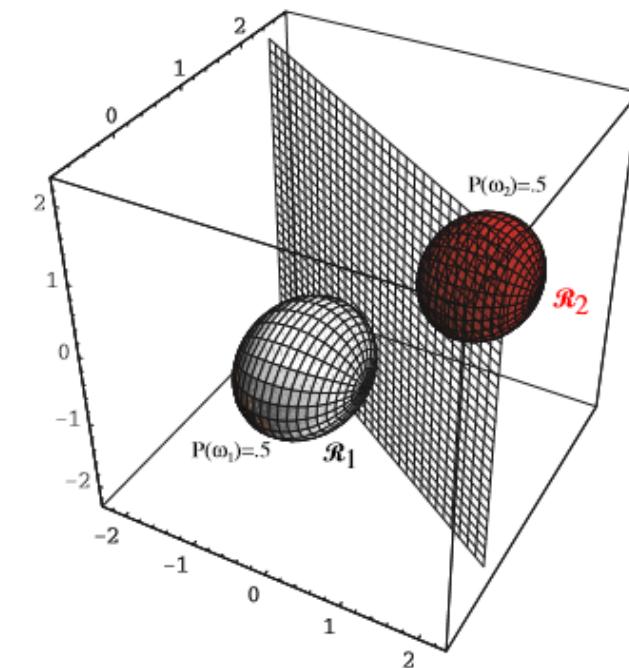
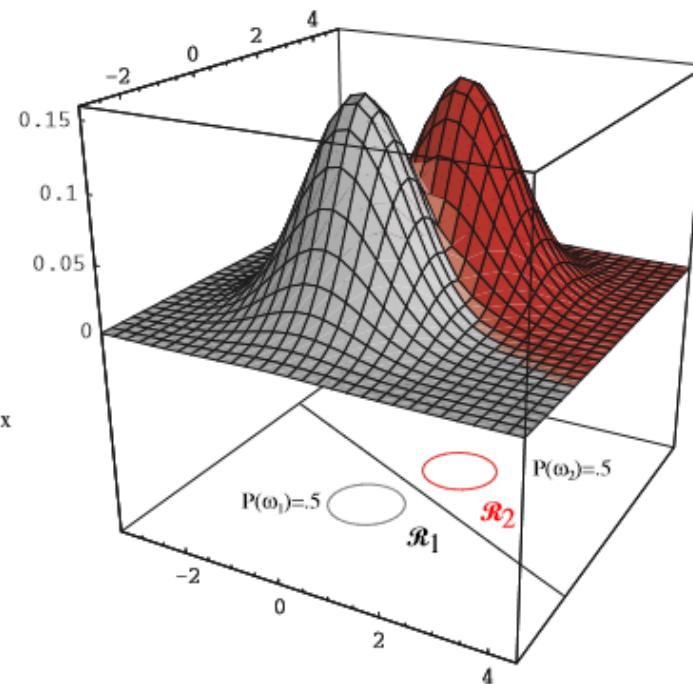
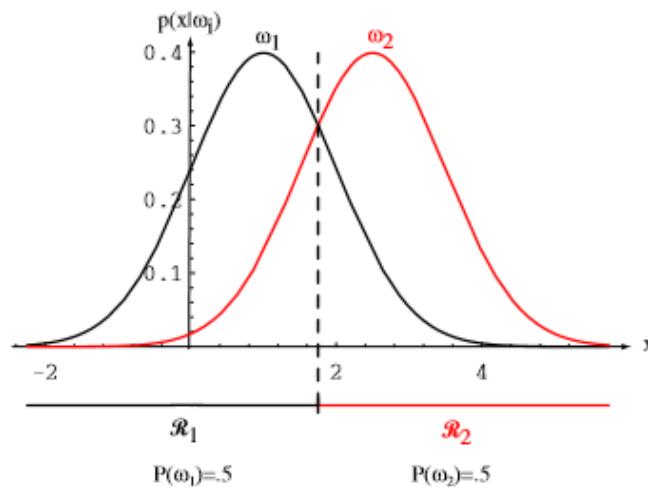
$$\begin{aligned} p(y) &= \phi^y(1-\phi)^{1-y} \\ p(x|y=0) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1} (x - \mu_0)\right) \\ p(x|y=1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1} (x - \mu_1)\right) \end{aligned}$$

For generative learning, the log likelihood is:

$$\begin{aligned} \ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^m p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) \\ &= \log \prod_{i=1}^m p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi) \end{aligned}$$



Higher Dimensional Gaussian



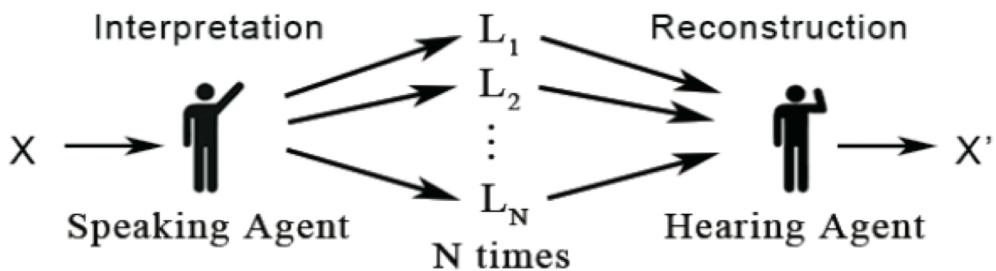
Q6: What the boundary looks like if we have different covariance matrix?

Semantics of Natural Language

Free Will

First published Mon Jan 7, 2002; substantive revision Fri Oct 29, 2010

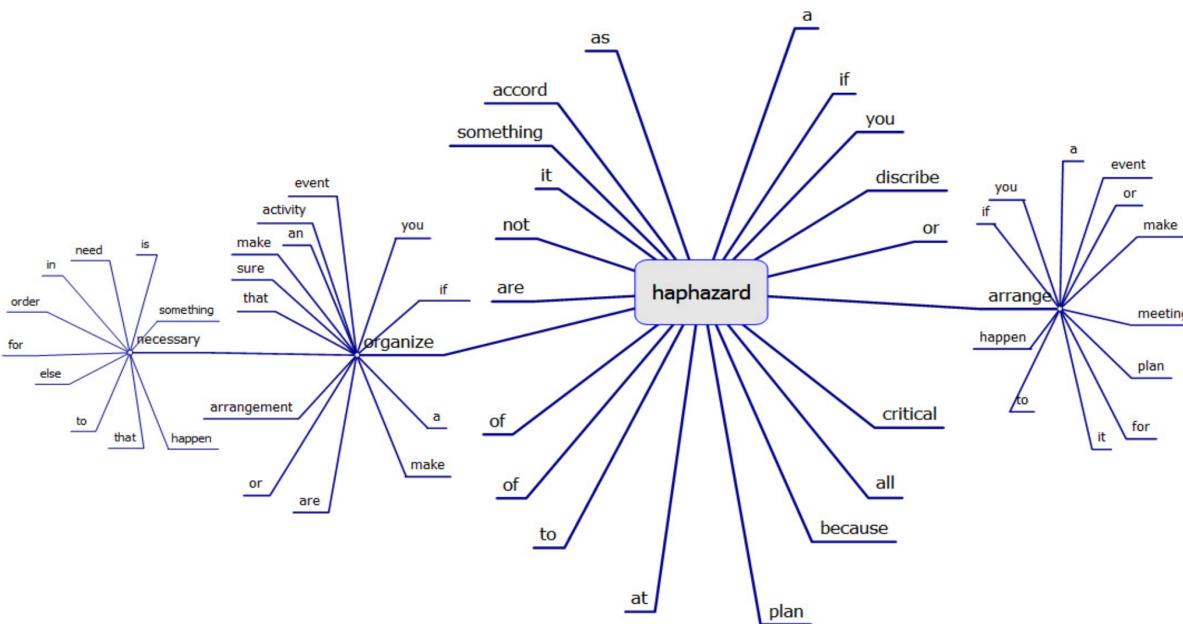
“Free Will” is a philosophical term of art for a particular sort of capacity of rational agents to choose a course of action from among various alternatives. Which sort is the free will sort is what all the fuss is about. (And what a fuss it has been: philosophers have debated this question for over two millennia, and just about every major philosopher has had something to say about it.) Most philosophers suppose that the concept of free will is very closely connected to the concept of moral responsibility. Acting with free will, on such views, is just to satisfy the metaphysical requirement on being responsible for one’s action. (Clearly, there will also be epistemic conditions on responsibility as well, such as being aware—or failing that, being culpably unaware—of relevant alternatives to one’s action and of the alternatives’ moral significance.) But the significance of free will is not exhausted by its connection to moral responsibility. Free will also appears to be a condition on desert for one’s accomplishments (why sustained effort and creative work are praiseworthy); on the autonomy and dignity of persons; and on the value we accord to love and friendship. (See Kane 1996, 81ff. and Clarke 2003, Ch. 1; but see also Pereboom 2001, Ch. 7.)



Bag-of-Features

- (1) John likes to watch movies. Mary likes movies too.
 - (2) John also likes to watch football games.

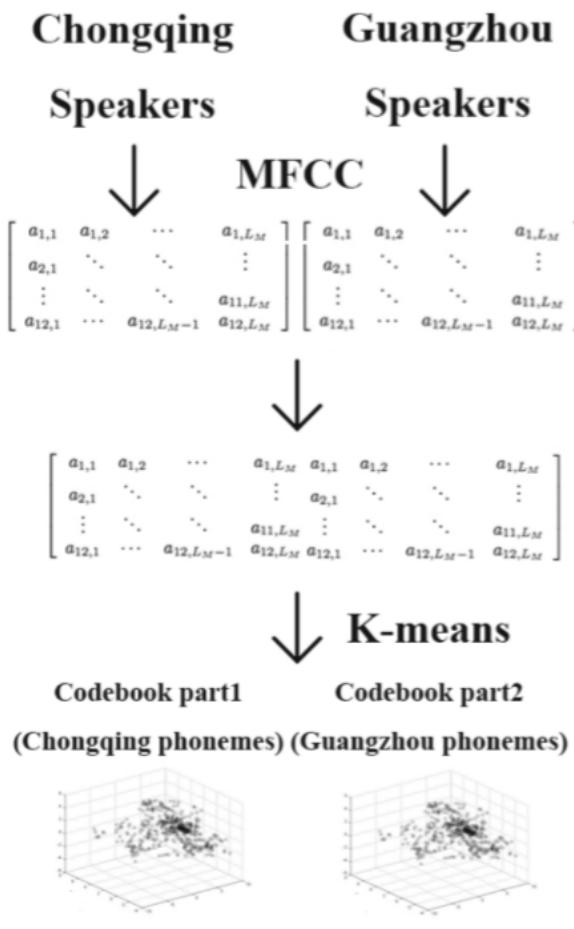
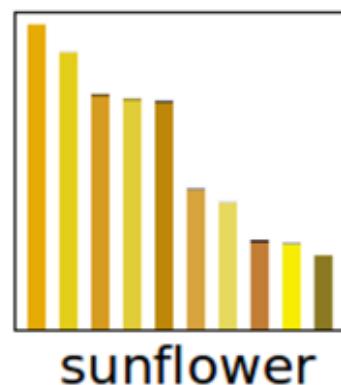
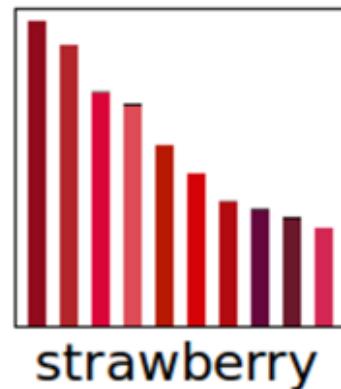
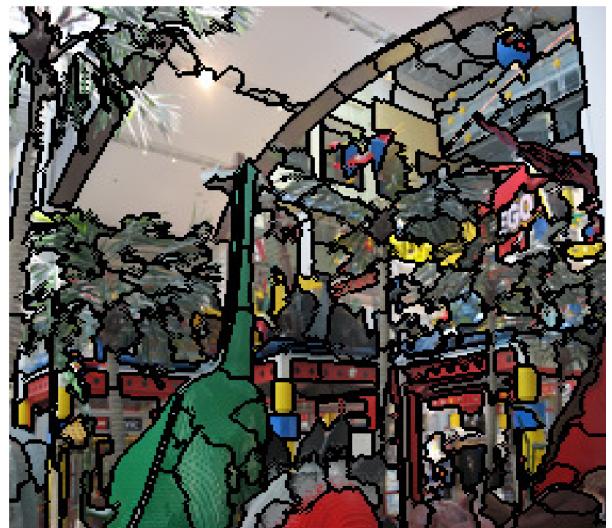
```
[ "John", "likes", "to", "watch",
  "movies", "Mary", "too", "also",
  "football", "games" ]
```



Minimum vocabulary construction – words are not all at the same level, we need to search down to get out these semantic bricks that build the whole language.



Bag-of-Features

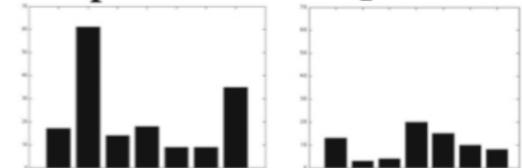


Codebook (all phonemes)

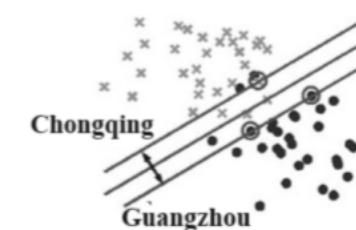
$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,L_M} & a_{1,1} & a_{1,2} & \cdots & a_{1,L_M} \\ a_{2,1} & \ddots & \ddots & \vdots & a_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{11,L_M} & \vdots & \ddots & \ddots & a_{11,L_M} \\ a_{12,1} & \cdots & a_{12,L_M-1} & a_{12,L_M} & a_{12,1} & \cdots & a_{12,L_M-1} & a_{12,L_M} \end{bmatrix}$$

Chongqing **Guangzhou**

Speaker **Speaker**



SVM Classifier

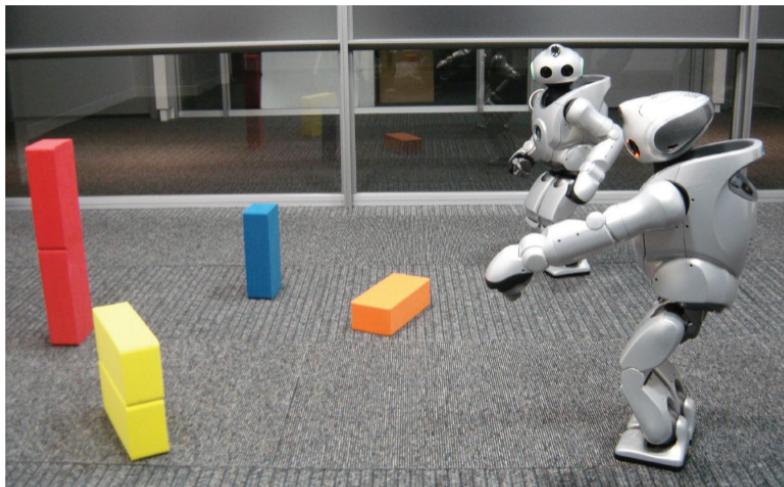
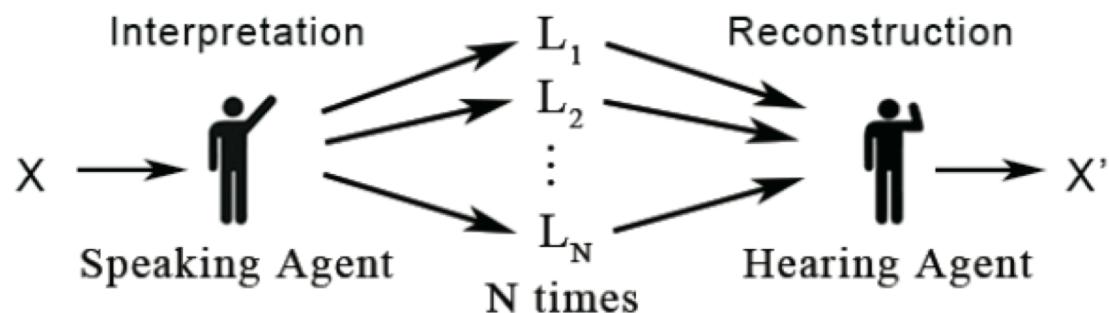


X. Zhang, Z. Qin, X. Liu and T. Wan: What color is an object? VCIP 2013: 1-5

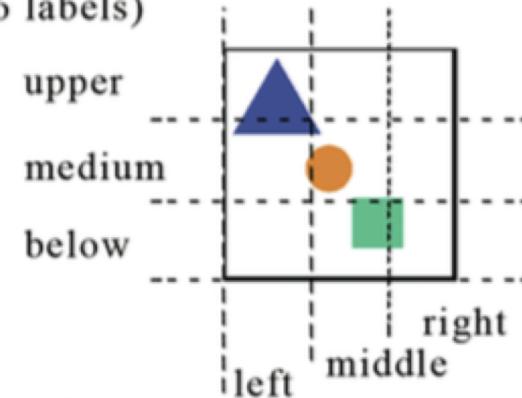
H. Zhao, Z. Qin, Y. Wang and Y. Wang :A Bag-of-phonemes Model for homeplace classification of Mandarin speakers, Iberian Conf. on Pattern Recognition and Image Analysis (IbPRIA 2015), LNCS 9117, pp. 683-690.

Understanding Concepts

Language Game: set of thought experiments that agents can understand each by interaction based on the physical world.



Position (6 labels)



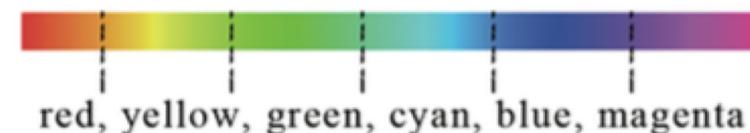
Shape (3 labels)



Size (3 labels)



Hue Color (6 labels)



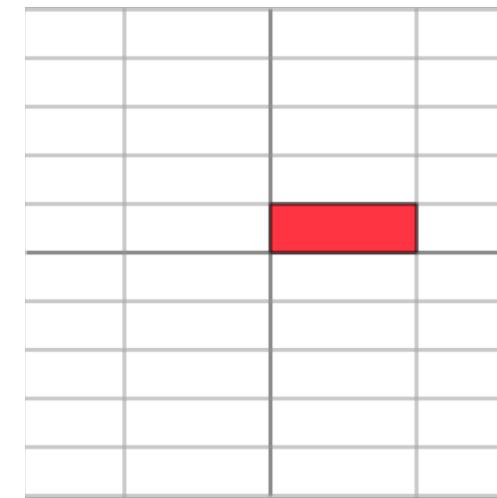
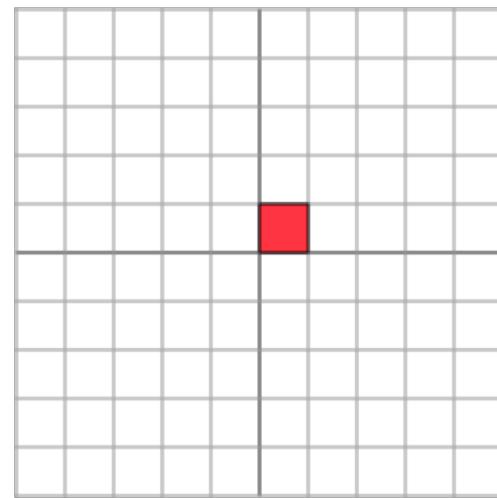
Latent Semantic Analysis

The geometry of linear transformations

Geometrically, we may think of a matrix like this as taking a point (x, y) in the plane and transforming it into another point using matrix multiplication:

$$M = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3x \\ y \end{bmatrix}.$$

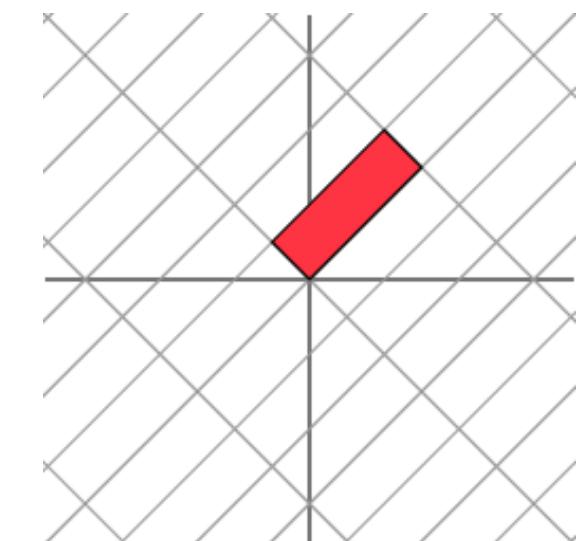
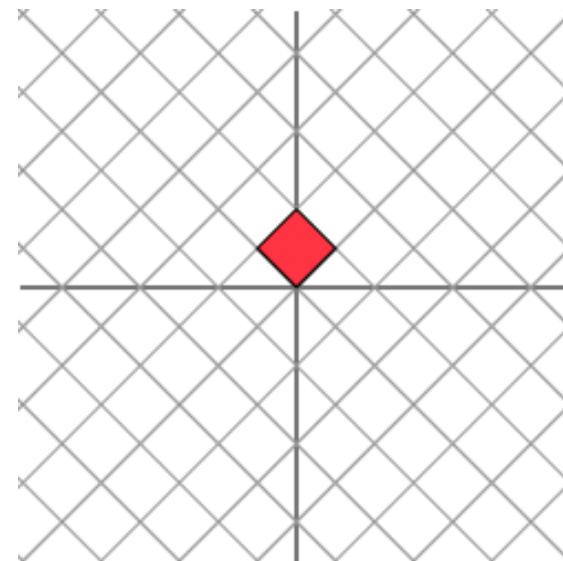
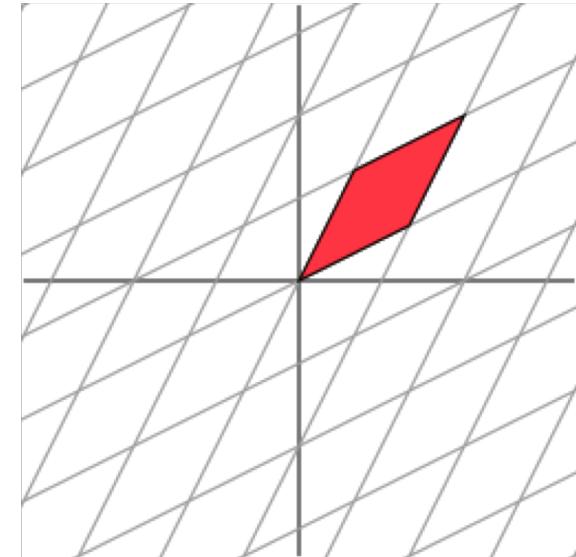
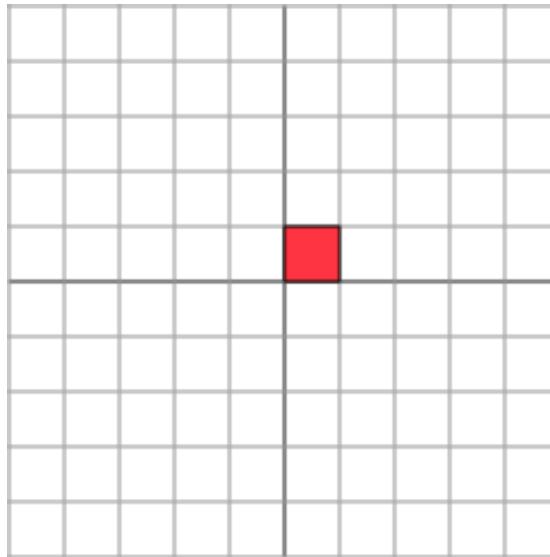


Another Example

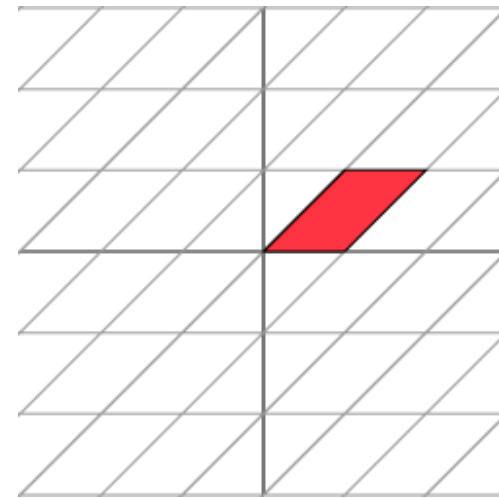
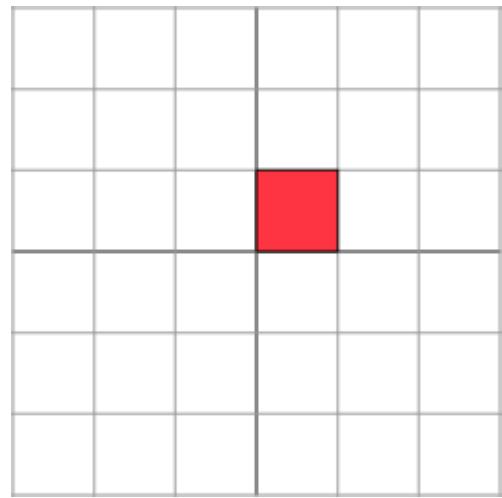
$$M = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

It is not so clear how to describe simply the geometric effect of the transformation. However, let's rotate our grid through a 45 degree angle and see what happens.

Symmetric matrices behave like diagonal matrices.

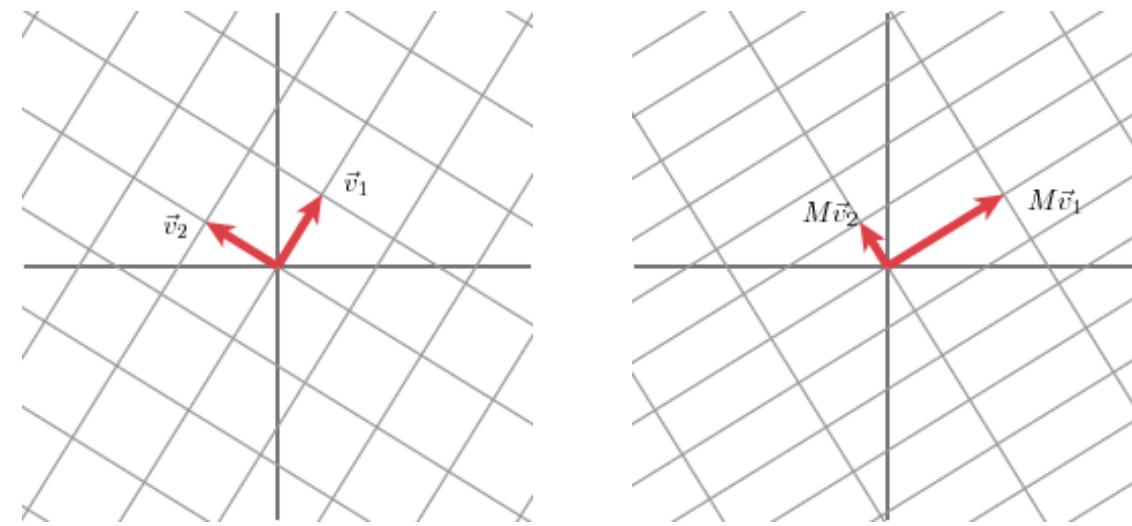


$$M = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$



The singular value decomposition

We will express this fact using vectors: with an appropriate choice of orthogonal unit vectors \mathbf{v}_1 and \mathbf{v}_2 , the vectors $M\mathbf{v}_1$ and $M\mathbf{v}_2$ are orthogonal.



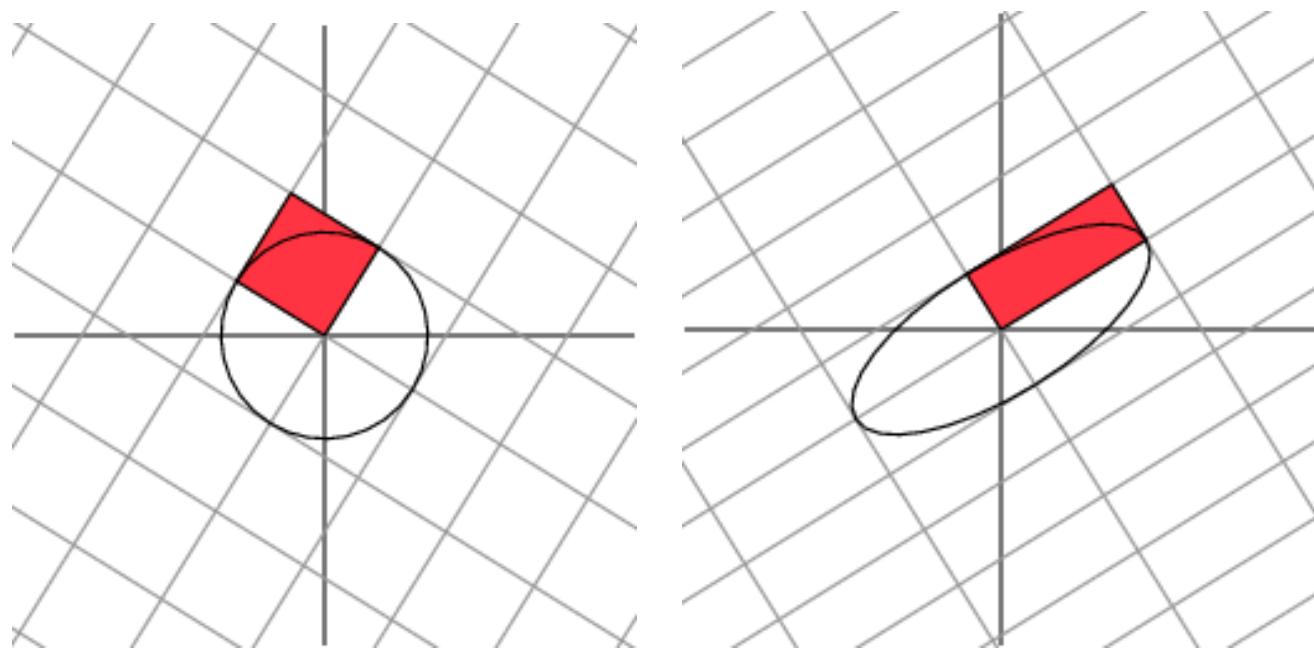
SVD

$$M = U\Sigma V^T$$

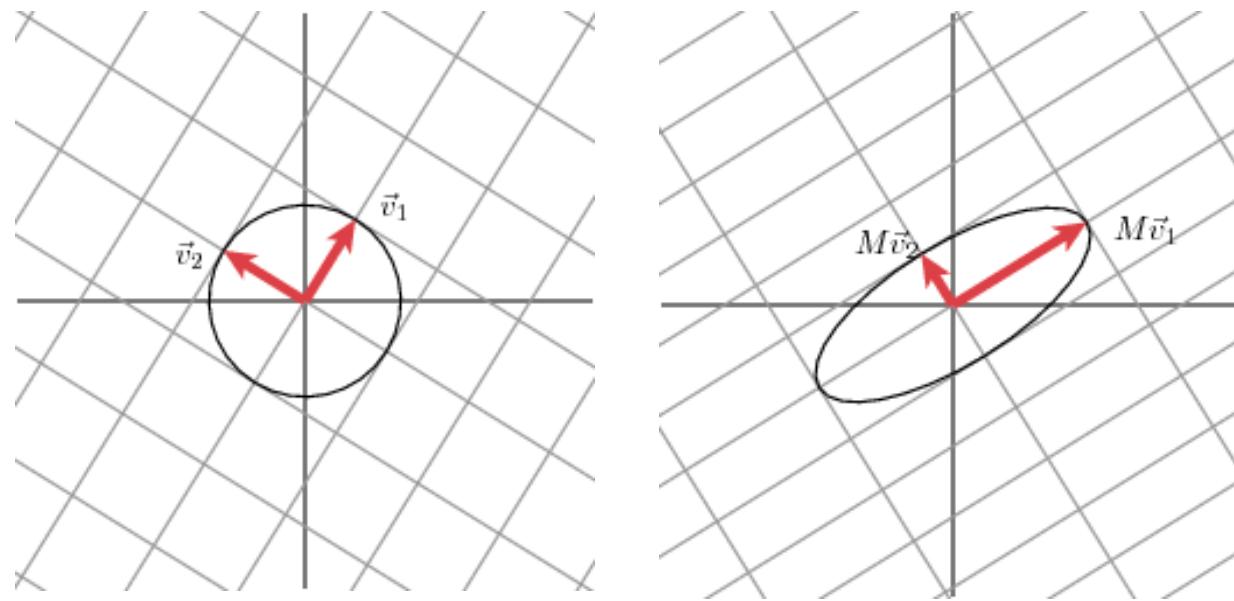
where U is a matrix whose columns are the vectors \mathbf{u}_1 and \mathbf{u}_2 , Σ is a diagonal matrix whose entries are σ_1 and σ_2 , and V is a matrix whose columns are \mathbf{v}_1 and \mathbf{v}_2 . The superscript T on the matrix V denotes the matrix transpose of V .

This shows how to decompose the matrix M into the product of three matrices: V describes an orthonormal basis in the domain, and U describes an orthonormal basis in the co-domain, and Σ describes how much the vectors in V are stretched to give the vectors in U .

Notice that the major and minor axes are defined by $M\mathbf{v}_1$ and $M\mathbf{v}_2$. These vectors therefore are the longest and shortest vectors among all the images of vectors on the unit circle.

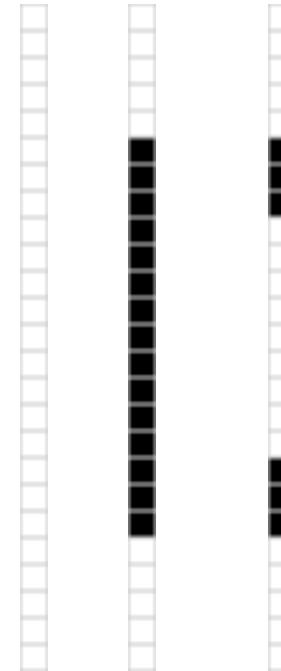
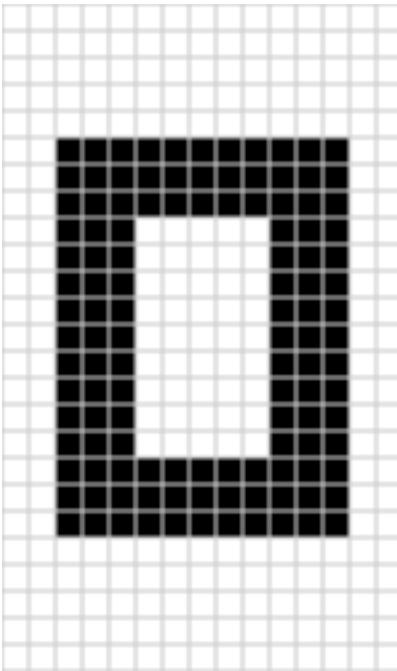


- A singular value decomposition provides a convenient way for breaking a matrix, which perhaps contains some data we are interested in, into simpler, meaningful pieces.

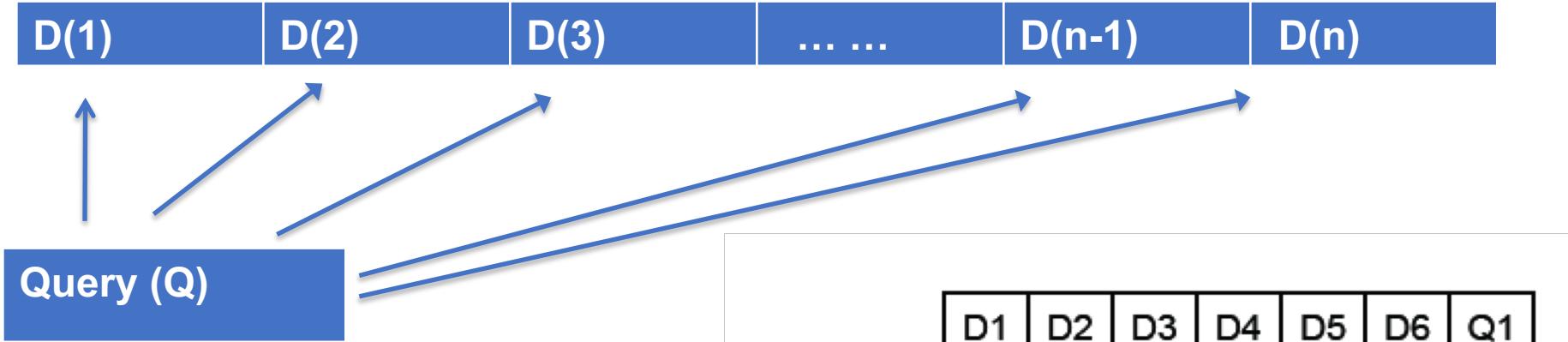


Data Compression

Singular value decompositions can be used to represent data efficiently.



Vector Representation



- Rock - (geology, music)?
- Context could help to solve the polysemy of words.
- Similarity: $\cos(x, y) = \frac{x \cdot y}{|x||y|}$.

	D1	D2	D3	D4	D5	D6	Q1
rock	2	1	0	2	0	1	1
granite	1	0	1	0	0	0	0
marble	1	2	0	0	0	0	1
music	0	0	0	1	2	0	0
song	0	0	0	1	0	2	0
band	0	0	0	0	1	0	0

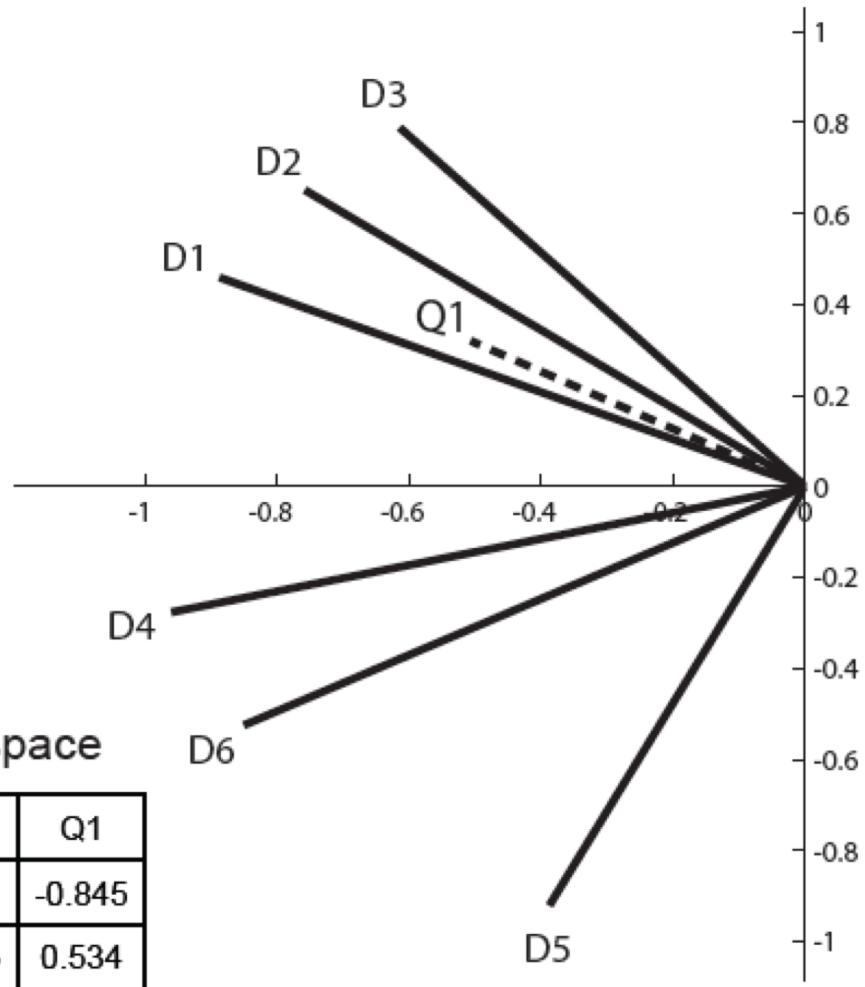
LSI – Latent Semantic Indexing

Original term-by-document matrix

	D1	D2	D3	D4	D5	D6	Q1
rock	2	1	0	2	0	1	1
granite	1	0	1	0	0	0	0
marble	1	2	0	0	0	0	1
music	0	0	0	1	2	0	0
song	0	0	0	1	0	2	0
band	0	0	0	0	1	0	0

Documents projected into 2D semantic space

	D1	D2	D3	D4	D5	D6	Q1
Dim. 1	-0.888	-0.759	-0.615	-0.961	-0.388	-0.851	-0.845
Dim. 2	0.460	0.652	0.789	-0.276	-0.922	-0.525	0.534



Topic Model

Probabilistic Latent Semantic Analysis

Probabilistic latent semantic analysis (PLSA), is a statistical technique for the analysis of two-mode and co-occurrence data. In effect, one can derive a low-dimensional representation of the observed variables in terms of their affinity to certain hidden variables, just as in latent semantic analysis, from which PLSA evolved.

Considering observations in the form of co-occurrences (w,d) of words and documents, PLSA models the probability of each co-occurrence as a mixture of conditionally independent multinomial distributions:

d is the document index variable, c is a word's topic drawn from the document's topic distribution, $P(c|d)$, and w is a word drawn from the word distribution of this word's topic, $P(w|c)$. The d and w are observable variables, the topic c is a latent variable.

$$P(w, d) = \sum_c P(c) P(d|c) P(w|c) = P(d) \sum_c P(c|d) P(w|c)$$

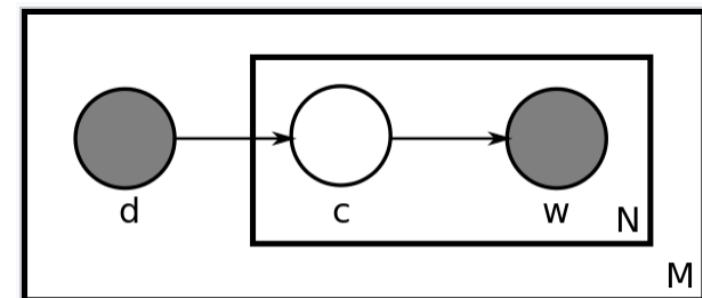
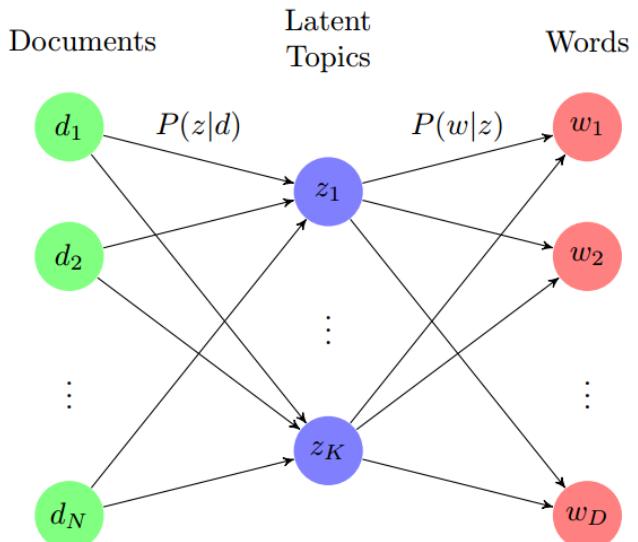


Plate Representation of P-LSA



The joint probability is represented by:

$$P(d, w) = P(d)P(w|d)$$

$$\begin{aligned} P(w|d) &= \sum_{z \in \mathcal{Z}} P(w, z|d) \\ &= \sum_{z \in \mathcal{Z}} P(w|d, z)P(z|d) \end{aligned}$$

$$P(w|d) = \sum_{z \in \mathcal{Z}} P(w|z)P(z|d)$$

$$P(w, d) = \sum_{z \in \mathcal{Z}} P(z)P(d|z)P(w|z)$$

The likelihood is written as:

$$L = \prod_{(d,w)} P(w|d) = \prod_{d \in \mathcal{D}} \prod_{w \in \mathcal{W}} P(w|d)^{n(d,w)}$$

$$\mathcal{L} = \log L = \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} n(d,w) \cdot \log \sum_{z \in \mathcal{Z}} P(w|z)P(z|d)$$

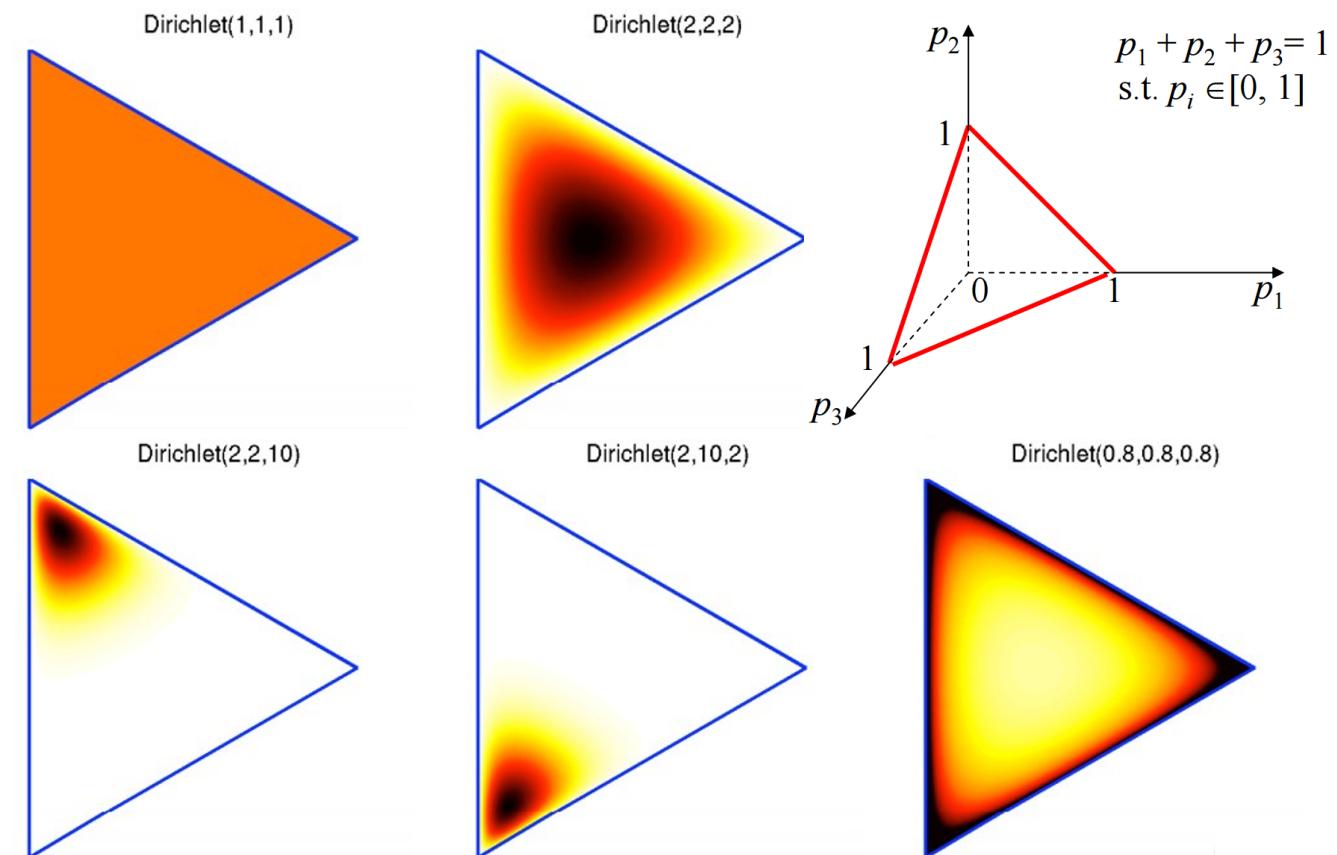
Dirichlet Distribution

The Dirichlet distribution is the multivariate generalization of the beta distribution. A Dirichlet distribution can be conceptualized as a probability distribution of distributions.

Mathematically, it is represented by:

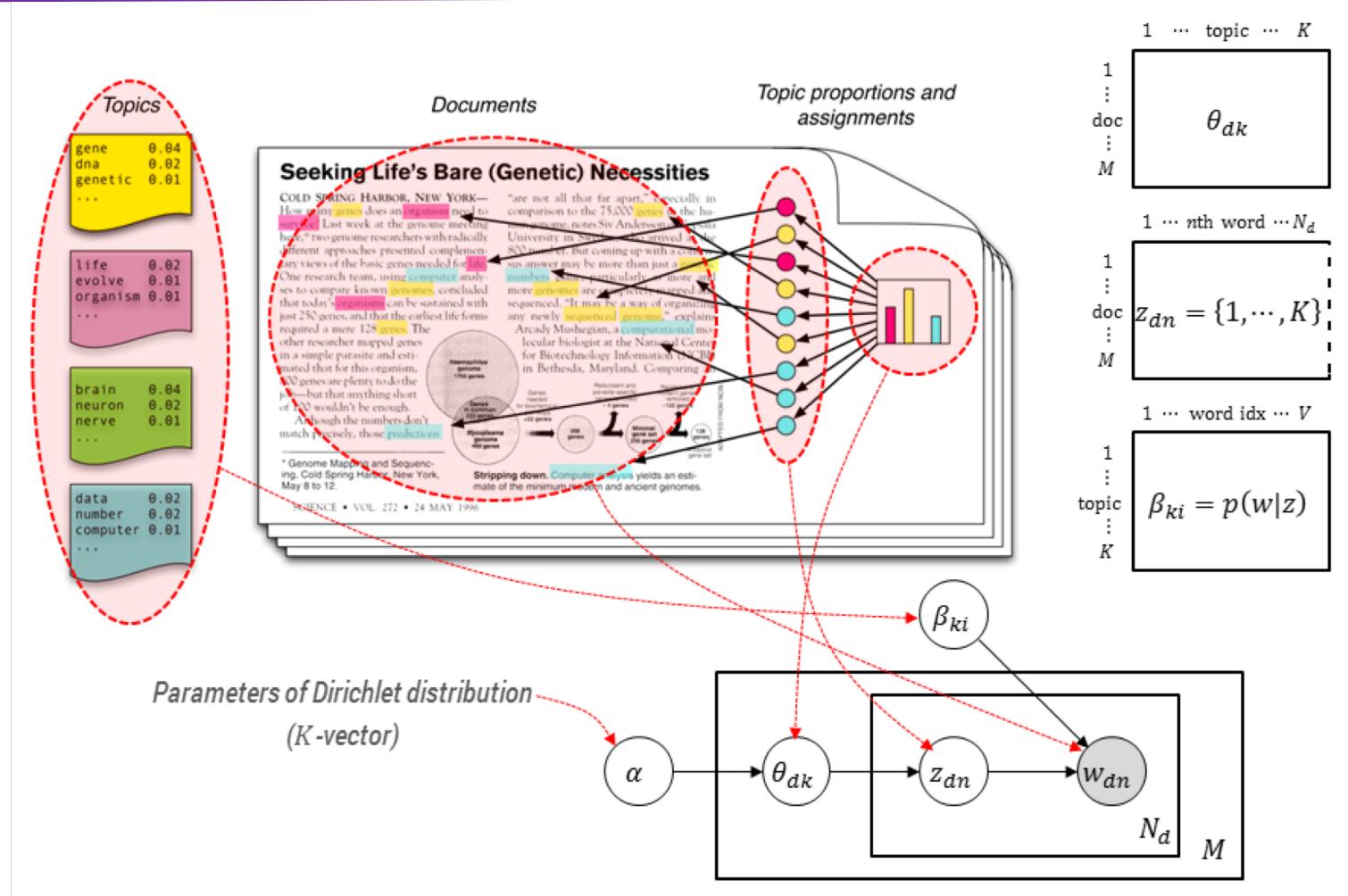
$$f(x_1, x_2, \dots, x_K; \alpha_1, \alpha_2, \dots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K x_i^{\alpha_i-1}$$

Where $0 < x_i < 1$ and $\sum_{i=1}^K x_i = 1$



Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a hierarchical generative statistical model. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.



An Example of Topics

Topic samples from training
(Top words of the list).

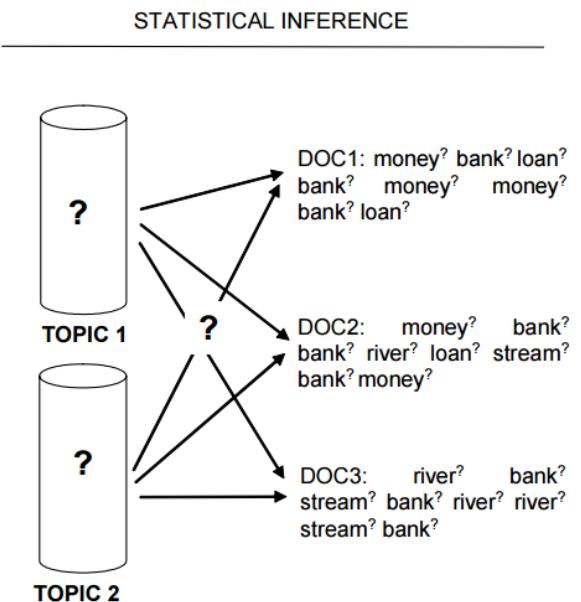
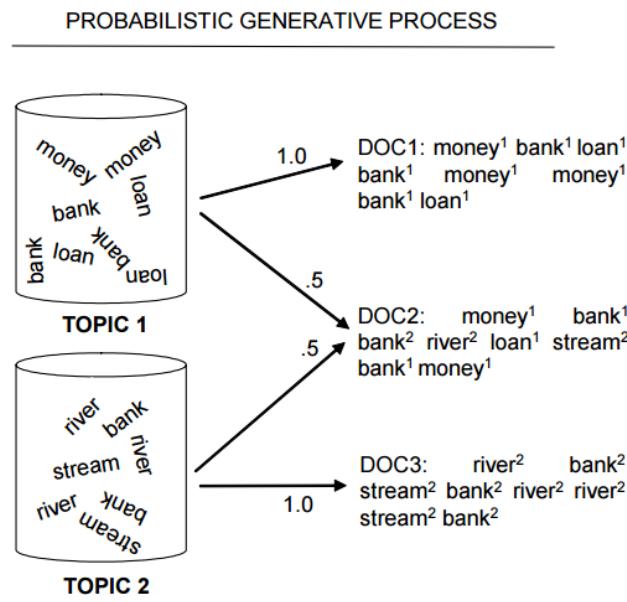
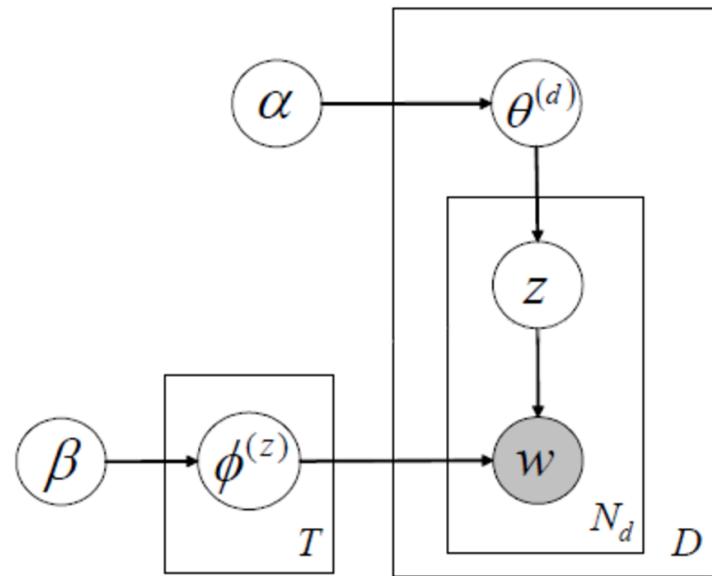
“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY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

Each topic is a distribution of words and a document can be represented as a mixture of topics. The following paragraph is a mixture of Arts, Budgets, Children and Education.

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.

Topic Models

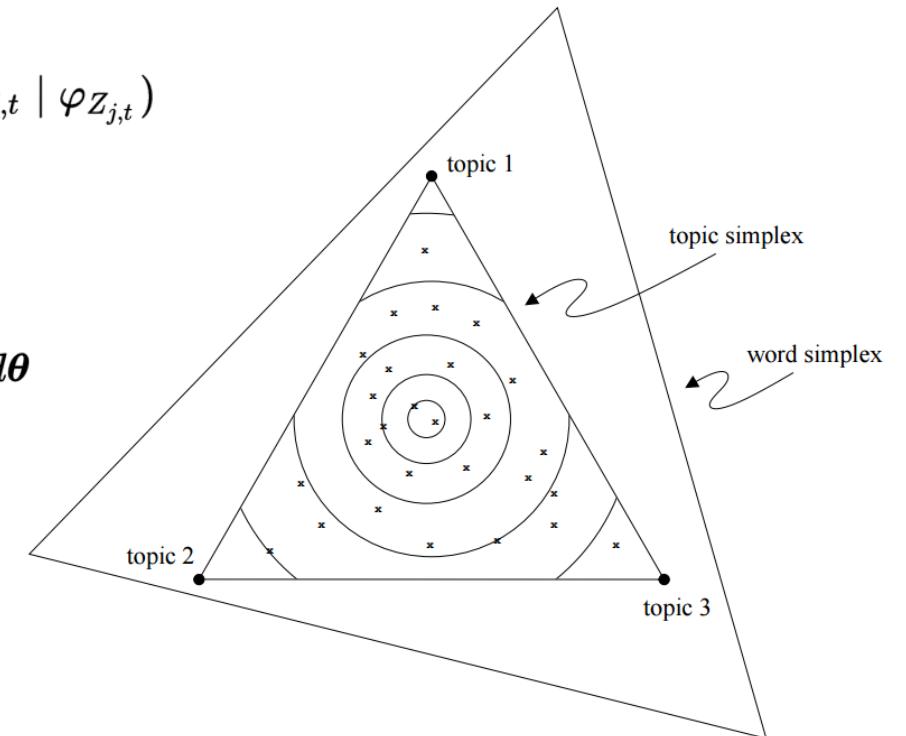
Illustration of the generative process and the problem of statistical inference underlying topic models



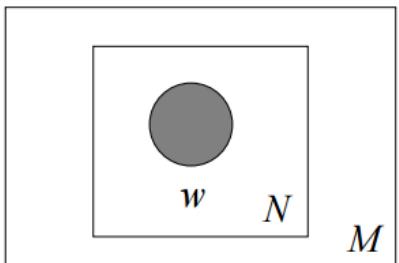
Likelihood

Illustration of the generative process and the problem of statistical inference underlying topic models

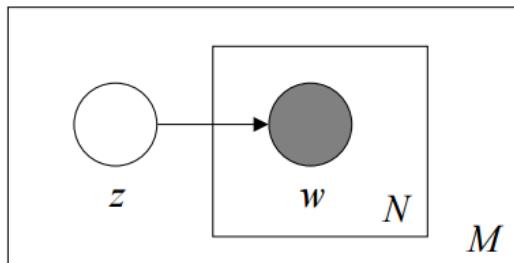
$$P(\mathbf{W}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{i=1}^K P(\varphi_i; \beta) \prod_{j=1}^M P(\theta_j; \alpha) \prod_{t=1}^N P(Z_{j,t} | \theta_j) P(W_{j,t} | \varphi_{Z_{j,t}})$$
$$P(\mathbf{Z}, \mathbf{W}; \alpha, \beta) = \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\varphi}} P(\mathbf{W}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) d\boldsymbol{\varphi} d\boldsymbol{\theta}$$
$$= \int_{\boldsymbol{\varphi}} \prod_{i=1}^K P(\varphi_i; \beta) \prod_{j=1}^M \prod_{t=1}^N P(W_{j,t} | \varphi_{Z_{j,t}}) d\boldsymbol{\varphi} \int_{\boldsymbol{\theta}} \prod_{j=1}^M P(\theta_j; \alpha) \prod_{t=1}^N P(Z_{j,t} | \theta_j) d\boldsymbol{\theta}$$



Model Comparisons



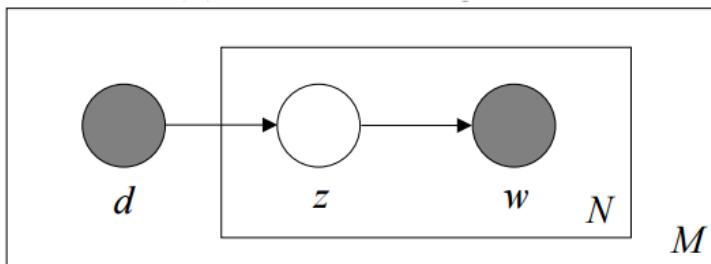
(a) unigram



(b) mixture of unigrams

$$p(\mathbf{w}) = \prod_{n=1}^N p(w_n)$$

$$p(\mathbf{w}) = \sum_z p(z) \prod_{n=1}^N p(w_n | z)$$

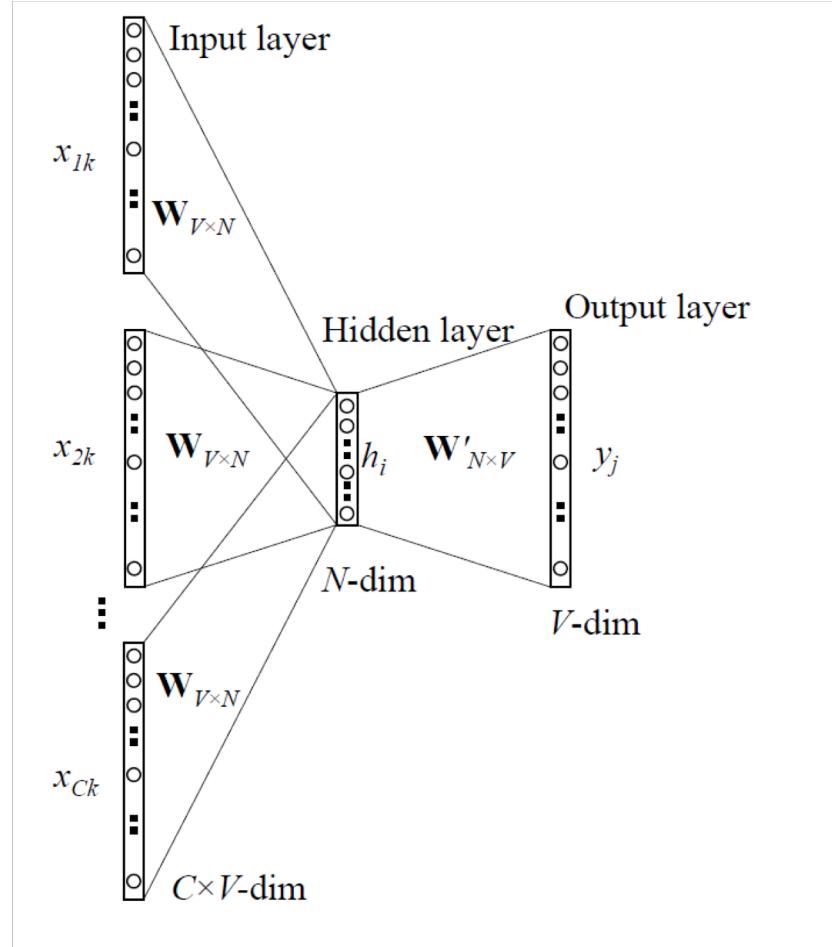


(c) pLSI/aspect model

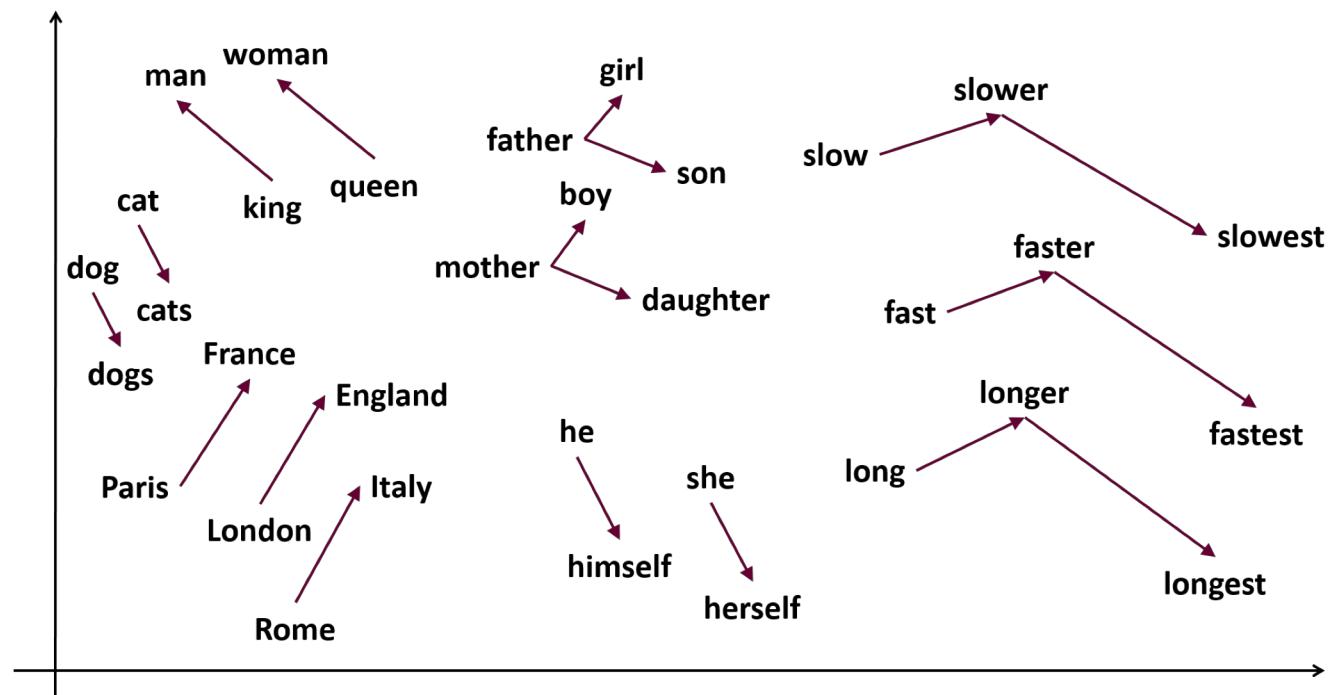
We compare LDA to simpler latent variable models for text—the unigram model, a mixture of unigrams, and the pLSI model. The pLSI model attempts to relax the simplifying assumption made in the mixture of unigrams model that each document is generated from only one topic.

$$p(d, w_n) = p(d) \sum_z p(w_n | z) p(z | d)$$

Word2Vec



Word2vec is a group of related models that are used to produce word embedding. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words.



References

Andrew Ng, Lectures notes of Machine Learning, CS229 at Stanford.

T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning, 2001, Springer.

C. Bishop, Pattern Recognition and Machine Learning, 2006, Springer.

T. Mitchell, Machine Learning, McGraw Hill.

https://en.wikipedia.org/wiki/Kernel_density_estimation

<http://setosa.io/ev/principal-component-analysis/>

<http://people.cs.pitt.edu/~iyad/PCA.pdf>

http://sebastianraschka.com/Articles/2014_pca_step_by_step.html