# ColumbiaX: Machine Learning

## Lecture 18

Prof. John Paisley

Department of Electrical Engineering
& Data Science Institute

Columbia University

# TOPIC MODELING

# Models for text data

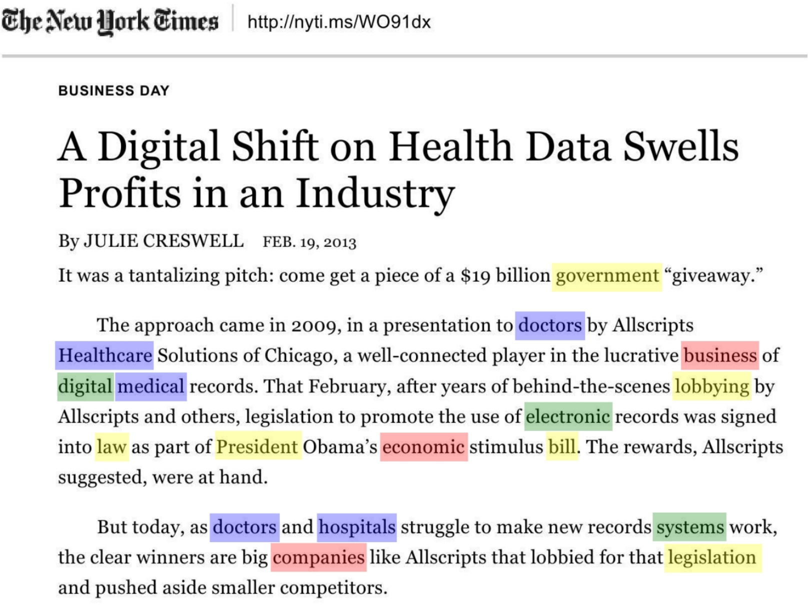
Given text data we want to:

- Organize
- Visualize
- Summarize
- Search
- Predict
- Understand

Topic models allow us to

1. Discover themes in text
2. Annotate documents
3. Organize, summarize, etc.

# TOPIC MODELING



The New York Times | http://nyti.ms/WO91dx

BUSINESS DAY

## A Digital Shift on Health Data Swells Profits in an Industry

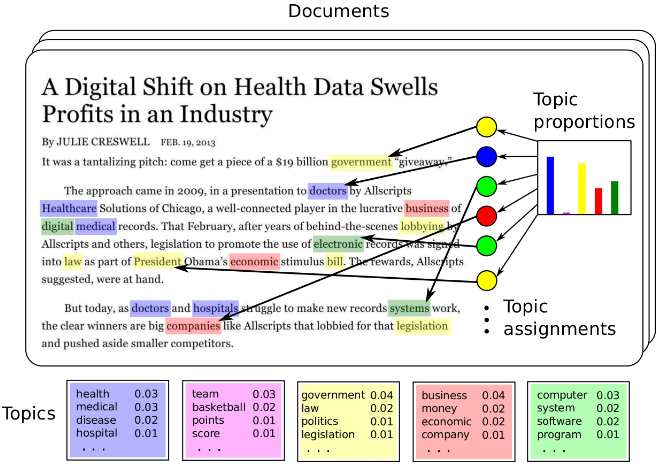
By JULIE CRESWELL FEB. 19, 2013

It was a tantalizing pitch: come get a piece of a $19 billion government "giveaway."

The approach came in 2009, in a presentation to doctors by Allscripts Healthcare Solutions of Chicago, a well-connected player in the lucrative business of digital medical records. That February, after years of behind-the-scenes lobbying by Allscripts and others, legislation to promote the use of electronic records was signed into law as part of President Obama's economic stimulus bill. The rewards, Allscripts suggested, were at hand.

But today, as doctors and hospitals struggle to make new records systems work, the clear winners are big companies like Allscripts that lobbied for that legislation and pushed aside smaller competitors.

# TOPIC MODELING



### A Digital Shift on Health Data Swells Profits in an Industry

By JULIE CRESWELL FEB. 19, 2013

It was a tantalizing pitch: come get a piece of a $19 billion government “giveaway.”

The approach came in 2009, in a presentation to doctors by Allscripts Healthcare Solutions of Chicago, a well-connected player in the lucrative business of digital medical records. That February, after years of behind-the-scenes lobbying by Allscripts and others, legislation to promote the use of electronic records was signed into law as part of President Obama’s economic stimulus bill. The rewards, Allscripts suggested, were at hand.

But today, as doctors and hospitals struggle to make new records systems work, the clear winners are big companies like Allscripts that lobbied for that legislation and pushed aside smaller competitors.

| | |
|---|---|
| health | 0.03 |
| medical | 0.03 |
| disease | 0.02 |
| hospital | 0.01 |
| . . . | |

| | |
|---|---|
| team | 0.03 |
| basketball | 0.02 |
| points | 0.01 |
| score | 0.01 |
| . . . | |

| | |
|---|---|
| government | 0.04 |
| law | 0.02 |
| politics | 0.01 |
| legislation | 0.01 |
| . . . | |

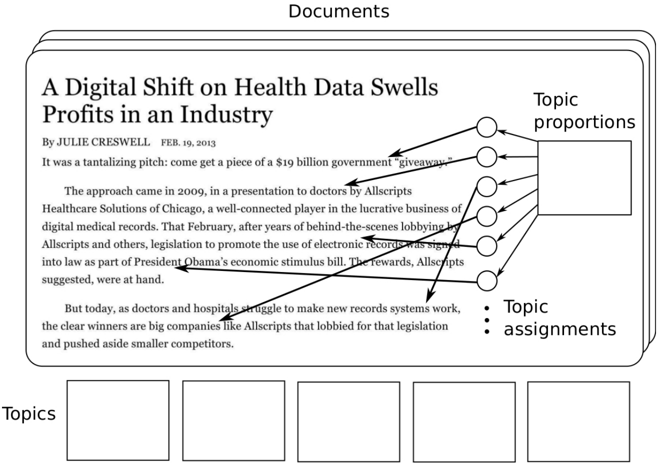
| | |
|---|---|
| business | 0.04 |
| money | 0.02 |
| economic | 0.02 |
| company | 0.01 |
| . . . | |

| | |
|---|---|
| computer | 0.03 |
| system | 0.02 |
| software | 0.02 |
| program | 0.01 |
| . . . | |

A probabilistic topic model

- Learns distributions on words called “topics” shared by documents
- Learns a distribution on topics for each document
- Assigns every word in a document to a topic

# TOPIC MODELING



A Digital Shift on Health Data Swells Profits in an Industry

By JULIE CRESWELL FEB. 19, 2013

It was a tantalizing pitch: come get a piece of a $19 billion government "giveaway."

The approach came in 2009, in a presentation to doctors by Allscripts Healthcare Solutions of Chicago, a well-connected player in the lucrative business of digital medical records. That February, after years of behind-the-scenes lobbying by Allscripts and others, legislation to promote the use of electronic records was signed into law as part of President Obama's economic stimulus bill. The rewards, Allscripts suggested, were at hand.

But today, as doctors and hospitals struggle to make new records systems work, the clear winners are big companies like Allscripts that lobbied for that legislation and pushed aside smaller competitors.

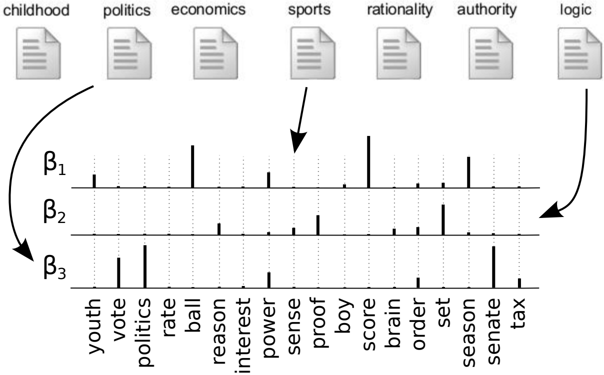
However, none of these things are known in advance and must be learned

- Each document is treated as a "bag of words"
- Need to define (1) a model, and (2) an algorithm to learn it
- We will review the standard topic model, but won't cover inference

# LATENT DIRICHLET ALLOCATION

There are two essential ingredients to latent Dirichlet allocation (LDA).

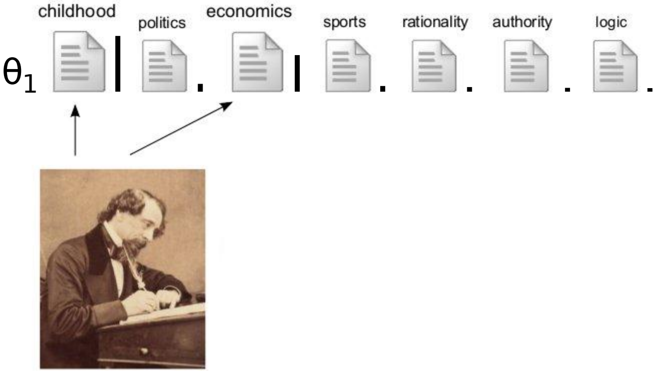
1. A collection of distributions on words (topics).
2. A distribution on topics for each document.



# LATENT DIRICHLET ALLOCATION

There are two essential ingredients to latent Dirichlet allocation (LDA).

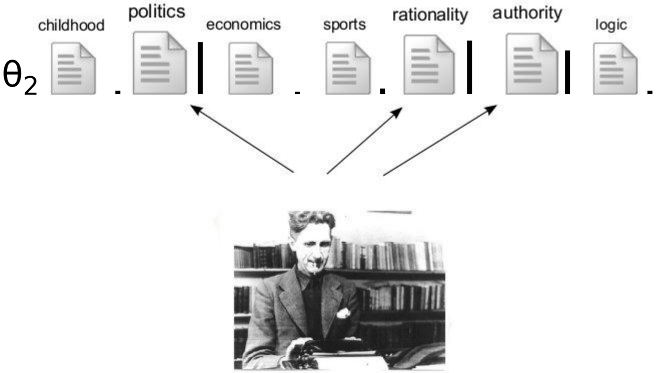
1. A collection of distributions on words (topics).
2. A distribution on topics for each document.



# Latent Dirichlet allocation

There are two essential ingredients to latent Dirichlet allocation (LDA).

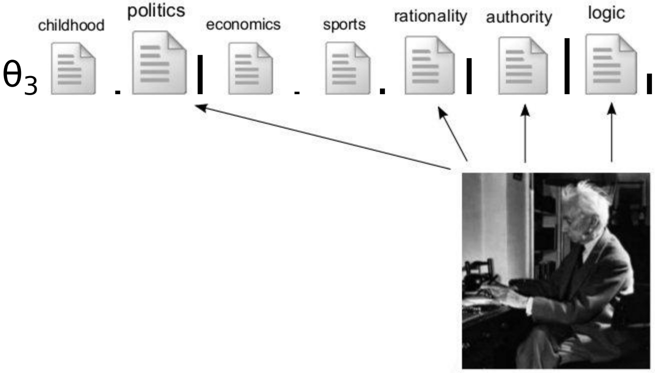
1. A collection of distributions on words (topics).
2. A distribution on topics for each document.



# LATENT DIRICHLET ALLOCATION

There are two essential ingredients to latent Dirichlet allocation (LDA).

1. A collection of distributions on words (topics).
2. A distribution on topics for each document.



# Latent Dirichlet allocation

There are two essential ingredients to latent Dirichlet allocation (LDA).

1. A collection of distributions on words (topics).
2. A distribution on topics for each document.

The generative process for LDA is:

1. Generate each topic, which is a distribution on words

$$\beta_k \sim \text{Dirichlet}(\gamma), \quad k = 1, \dots, K$$

2. For each document, generate a distribution on topics

$$\theta_d \sim \text{Dirichlet}(\alpha), \quad d = 1, \dots, D$$

3. For the *n*th word in the *d*th document,
   a) Allocate the word to a topic, $c_{dn} \sim \text{Discrete}(\theta_d)$
   b) Generate the word from the selected topic, $x_{dn} \sim \text{Discrete}(\beta_{c_{dn}})$

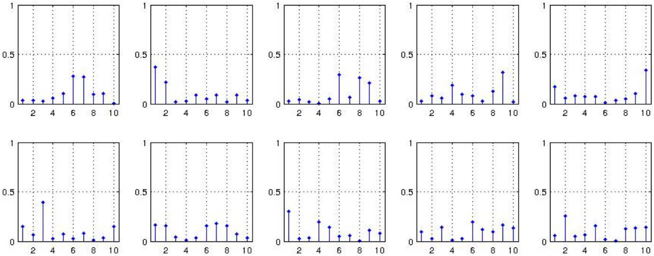
# DIRICHLET DISTRIBUTION

A continuous distribution on discrete probability vectors. Let $\beta_k$ be a probability vector and $\gamma$ a positive parameter vector,

$$p(\beta_k|\gamma) = \frac{\Gamma(\sum_v \gamma_v)}{\prod_{v=1}^{V} \Gamma(\gamma_v)} \prod_{v=1}^{V} \beta_{k,v}^{\gamma_v - 1}$$

This defines the Dirichlet distribution. Some examples of $\beta_k$ generated from this distribution for a constant value of $\gamma$ and $V = 10$ are given below.

$\gamma = 1$



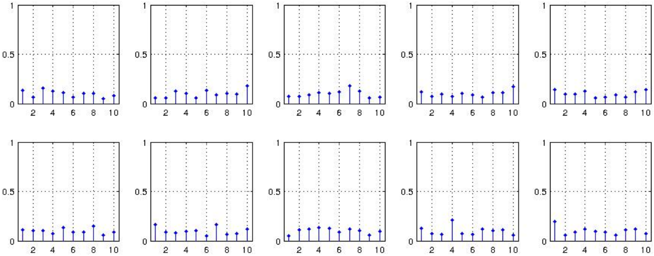
# DIRICHLET DISTRIBUTION

A continuous distribution on discrete probability vectors. Let $\beta_k$ be a probability vector and $\gamma$ a positive parameter vector,

$$p(\beta_k|\gamma) = \frac{\Gamma(\sum_v \gamma_v)}{\prod_{v=1}^V \Gamma(\gamma_v)} \prod_{v=1}^{V} \beta_{k,v}^{\gamma_v - 1}$$

This defines the Dirichlet distribution. Some examples of $\beta_k$ generated from this distribution for a constant value of $\gamma$ and $V = 10$ are given below.

$\gamma = 10$



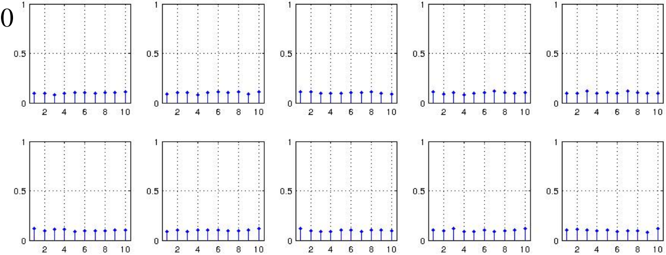
# DIRICHLET DISTRIBUTION

A continuous distribution on discrete probability vectors. Let $\beta_k$ be a probability vector and $\gamma$ a positive parameter vector,

$$p(\beta_k|\gamma) = \frac{\Gamma(\sum_v \gamma_v)}{\prod_{v=1}^{V} \Gamma(\gamma_v)} \prod_{v=1}^{V} \beta_{k,v}^{\gamma_v - 1}$$

This defines the Dirichlet distribution. Some examples of $\beta_k$ generated from this distribution for a constant value of $\gamma$ and $V = 10$ are given below.

$\gamma = 100$



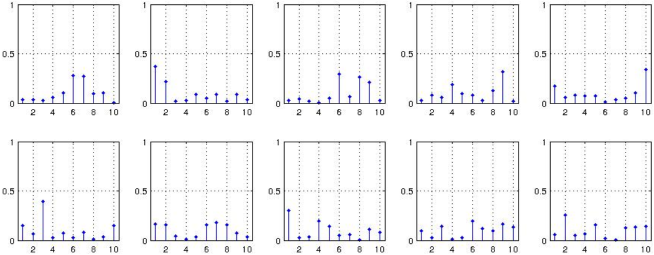
# DIRICHLET DISTRIBUTION

A continuous distribution on discrete probability vectors. Let $\beta_k$ be a probability vector and $\gamma$ a positive parameter vector,

$$p(\beta_k|\gamma) = \frac{\Gamma(\sum_v \gamma_v)}{\prod_{v=1}^{V}\Gamma(\gamma_v)} \prod_{v=1}^{V} \beta_{k,v}^{\gamma_v - 1}$$

This defines the Dirichlet distribution. Some examples of $\beta_k$ generated from this distribution for a constant value of $\gamma$ and $V = 10$ are given below.

$\gamma = 1$



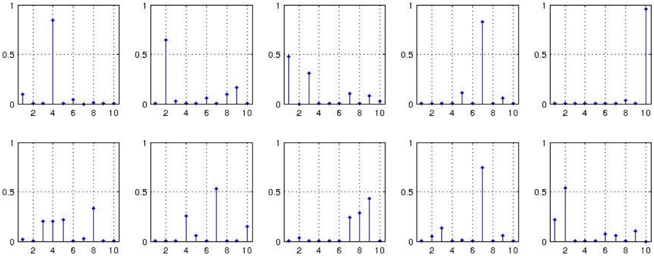
# DIRICHLET DISTRIBUTION

A continuous distribution on discrete probability vectors. Let $\beta_k$ be a probability vector and $\gamma$ a positive parameter vector,

$$p(\beta_k|\gamma) = \frac{\Gamma(\sum_v \gamma_v)}{\prod_{v=1}^{V}\Gamma(\gamma_v)} \prod_{v=1}^{V} \beta_{k,v}^{\gamma_v - 1}$$

This defines the Dirichlet distribution. Some examples of $\beta_k$ generated from this distribution for a constant value of $\gamma$ and $V = 10$ are given below.

$\gamma = 0.1$



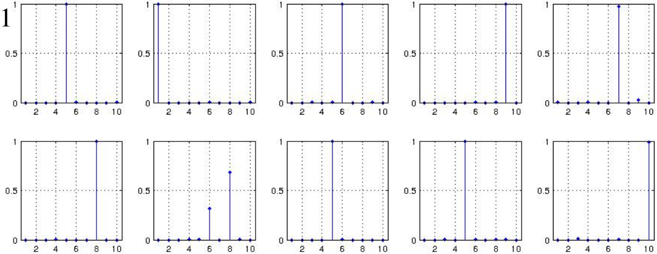
# DIRICHLET DISTRIBUTION

A continuous distribution on discrete probability vectors. Let $\beta_k$ be a probability vector and $\gamma$ a positive parameter vector,

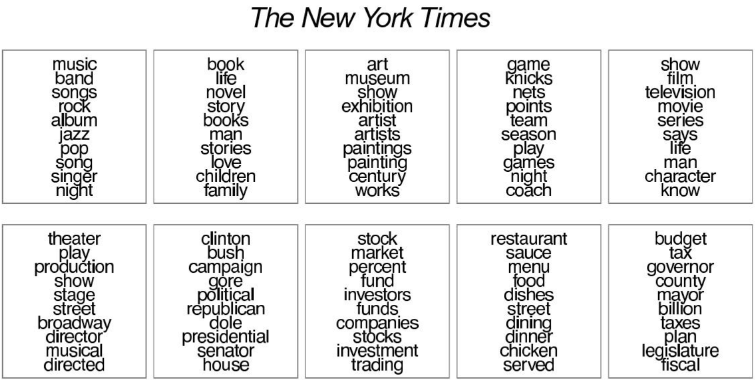
$$p(\beta_k|\gamma) = \frac{\Gamma(\sum_v \gamma_v)}{\prod_{v=1}^{V} \Gamma(\gamma_v)} \prod_{v=1}^{V} \beta_{k,v}^{\gamma_v - 1}$$

This defines the Dirichlet distribution. Some examples of $\beta_k$ generated from this distribution for a constant value of $\gamma$ and $V = 10$ are given below.

$\gamma = 0.01$



# LDA OUTPUT



*The New York Times*

| | | | | |
|---|---|---|---|---|
| music | book | art | game | show |
| band | life | museum | knicks | film |
| songs | novel | show | nets | television |
| rock | story | exhibition | points | movie |
| album | books | artist | team | series |
| jazz | man | artists | season | says |
| pop | stories | paintings | play | life |
| song | love | painting | games | man |
| singer | children | century | night | character |
| night | family | works | coach | know |

| | | | | |
|---|---|---|---|---|
| theater | clinton | stock | restaurant | budget |
| play | bush | market | sauce | tax |
| production | campaign | percent | menu | governor |
| show | gore | fund | food | county |
| stage | political | investors | dishes | mayor |
| street | republican | funds | street | billion |
| broadway | dole | companies | dining | taxes |
| director | presidential | stocks | dinner | plan |
| musical | senator | investment | chicken | legislature |
| directed | house | trading | served | fiscal |

LDA outputs two main things:

1. A set of distributions on words (topics). Shown above are ten topics from NYT data. We list the ten words with the highest probability.
2. A distribution on topics for each document (not shown). This indicates its thematic breakdown and provides a compact representation.

# LDA and Matrix Factorization

**Q**: For a particular document, what is $P(x_{dn} = i|\boldsymbol{\beta}, \theta_d)$?

**A**: Find this by integrating out the cluster assignment,

$$\begin{aligned} P(x_{dn} = i|\boldsymbol{\beta}, \theta) &= \sum_{k=1}^{K} P(x_{dn} = i, c_{dn} = k|\boldsymbol{\beta}, \theta_d) \\ &= \sum_{k=1}^{K} \underbrace{P(x_{dn} = i, |\boldsymbol{\beta}, c_{dn} = k)}_{= \beta_{ki}} \underbrace{P(c_{dn} = k|\theta_d)}_{= \theta_{dk}} \end{aligned}$$

Let $B = [\beta_1, \ldots, \beta_K]$ and $\Theta = [\theta_1, \ldots, \theta_D]$, then $P(x_{dn} = i|\beta, \theta) = (B\Theta)_{id}$

In other words, we can read the probabilities from a matrix formed by taking the product of two matrices that have nonnegative entries.

# NONNEGATIVE MATRIX FACTORIZATION

# NONNEGATIVE MATRIX FACTORIZATION

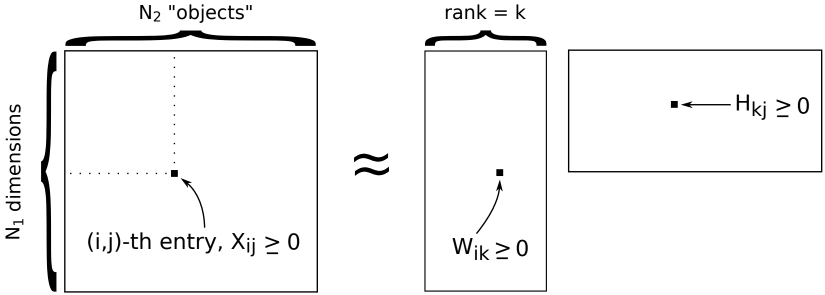
LDA can be thought of as an instance of nonnegative matrix factorization.

- It is a probabilistic model.
- Inference involves techniques not taught in this course.

We will discuss two other related models and their algorithms. These two models are called *nonnegative matrix factorization* (NMF)

- They can be used for the same tasks as LDA
- Though "nonnegative matrix factorization" is a general technique, "NMF" usually just refers to the following two methods.

# Nonnegative matrix factorization



We use notation and think about the problem slightly differently from PMF

- Data $X$ has nonnegative entries. None missing, but likely many zeros.
- The learned factorization $W$ and $H$ also have nonnegative entries.
- The value $X_{ij} \approx \sum_k W_{ik}H_{kj}$, but we won't write this with vector notation
- Later we interpret the output in terms of columns of $W$ and $H$.

# Nonnegative matrix factorization

What are some data modeling problems that can constitute $X$?

- Text data:
  - Word term frequencies
  - $X_{ij}$ contains the number of times word $i$ appears in document $j$.
- Image data:
  - Face identification data sets
  - Put each *vectorized* $N \times M$ image of a face on a *column* of $X$.
- Other discrete grouped data:
  - Quantize *continuous* sets of features using K-means
  - $X_{ij}$ counts how many times group $j$ uses cluster $i$.
  - For example: group = song, features = $d \times n$ spectral information matrix

# Two objective functions

NMF minimizes one of the following two objective functions over $W$ and $H$.

**Choice 1: Squared error objective**

$$\|X - WH\|^2 = \sum_i \sum_j (X_{ij} - (WH)_{ij})^2$$

**Choice 2: Divergence objective**

$$D(X\|WH) = -\sum_i \sum_j \left[X_{ij} \ln(WH)_{ij} - (WH)_{ij}\right]$$

- Both have the constraint that $W$ and $H$ contain nonnegative values.
- NMF uses a fast, simple algorithm for optimizing these two objectives.

# MINIMIZATION AND MULTIPLICATIVE ALGORITHMS[1]

Recall what we should look for in minimizing an objective "$\min_h F(h)$":

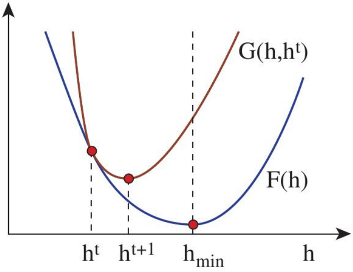
1. A way to generate a sequence of values $h^1, h^2, \ldots$, such that

$$F(h^1) \geq F(h^2) \geq F(h^3) \geq \cdots$$

2. Convergence of the sequence to a local minimum of $F$

The following algorithms fulfill these requirements. In this case:

- Minimization is done via an "auxiliary function."
- Leads to a "multiplicative algorithm" for $W$ and $H$.
- We'll skip details (see reference).



[1] For details, see D.D. Lee and H.S. Seung (2001). "Algorithms for non-negative matrix factorization." *Advances in Neural Information Processing Systems.*

# Multiplicative update for $\|X - WH\|^2$

**Problem**

$$\min \ \textstyle\sum_{ij}(X_{ij} - (WH)_{ij})^2 \qquad \text{subject to } \ W_{ik} \geq 0, \ H_{kj} \geq 0.$$

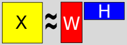
**Algorithm**

- Randomly initialize $H$ and $W$ with nonnegative values.
- Iterate the following, first for all values in $H$, then all in $W$:

$$\begin{aligned} H_{kj} &\leftarrow H_{kj}\frac{(W^T X)_{kj}}{(W^T WH)_{kj}}, \\ W_{ik} &\leftarrow W_{ik}\frac{(XH^T)_{ik}}{(WHH^T)_{ik}}, \end{aligned}$$

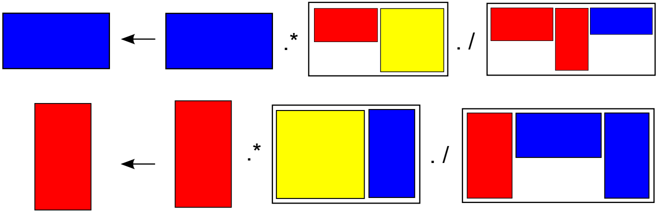
until the change in $\|X - WH\|^2$ is “small.”

# Visualization and maximum likelihood



A visualization that may be helpful. Use the color-coded definition above.

- Use element-wise multiplication/division across three columns below.
- Use matrix multiplication within each outlined box.



Probabilistically, the squared error penalty implies a Gaussian distribution,

$$X_{ij} \sim N(\textstyle\sum_k W_{ik}H_{kj}, \sigma^2)$$

Since $X_{ij} \geq 0$ (and often isn't continuous), we are making an incorrect modeling assumption. Nevertheless, as with PMF it still works well.

# Multiplicative update for $D(X\|WH)$

**Problem**

$$\min \sum_{ij} \left[X_{ij} \ln \tfrac{1}{(WH)_{ij}} + (WH)_{ij}\right] \quad \text{subject to } W_{ik} \geq 0,\ H_{kj} \geq 0.$$

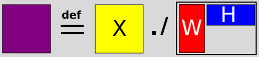
**Algorithm**

- Randomly initialize $H$ and $W$ with nonnegative values.
- Iterate the following, first for all values in $H$, then all in $W$:

$$\begin{aligned} H_{kj} &\leftarrow H_{kj} \frac{\sum_i W_{ik} X_{ij}/(WH)_{ij}}{\sum_i W_{ik}}, \\ W_{ik} &\leftarrow W_{ik} \frac{\sum_j H_{kj} X_{ij}/(WH)_{ij}}{\sum_j H_{kj}}, \end{aligned}$$

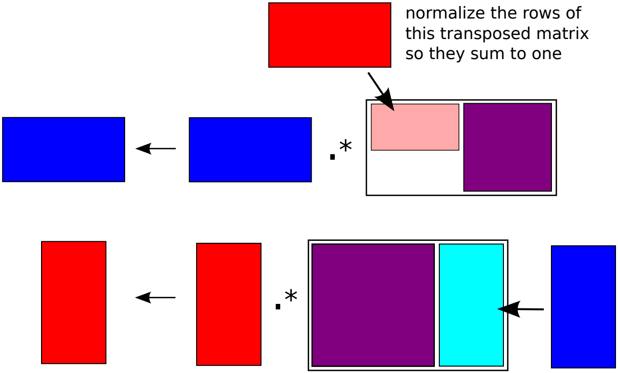
until the change in $D(X\|WH)$ is “small.”

# VISUALIZATION



Visualizing the update for the divergence penalty is more complicated.

- Use the color-coded definition above.
- "Purple" is the data matrix "dot-divided" by the approximation of it.



normalize the rows of
this transposed matrix
so they sum to one

normalize the columns
of this matrix so they
sum to one

# MAXIMUM LIKELIHOOD

The maximum likelihood interpretation of the divergence penalty is more interesting than for the squared error penalty.

If we model the data as independent Poisson random variables

$$X_{ij} \sim \text{Pois}((WH)_{ij}), \qquad \text{Pois}(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \;\; x \in \{0, 1, 2, \dots\},$$

then the negative divergence penalty is maximum likelihood for $W$ and $H$.

$$\begin{aligned} -D(X\|WH) &= \sum_{ij} [X_{ij} \ln(WH)_{ij} - (WH)_{ij}] \\ &= \sum_{ij} \ln P(X_{ij}|W, H) + \text{constant} \end{aligned}$$

We use: $P(X|W, H) = \prod_{ij} P(X_{ij}|W, H) = \prod_{ij} \text{Pois}(X_{ij}|(WH)_{ij})$.

# NMF AND TOPIC MODELING

As discussed, NMF can be used for topic modeling. In fact, one can show that the divergence penalty is closely related mathematically to LDA.

Step 1. Form the term-frequency matrix $X$. ($X_{ij} = \#$ times word $i$ in doc $j$)

Step 2. Run NMF to learn $W$ and $H$ using $D(X\|WH)$ penalty

Step 3. As an added step, after Step 2 is complete, for $k = 1, \dots, K$

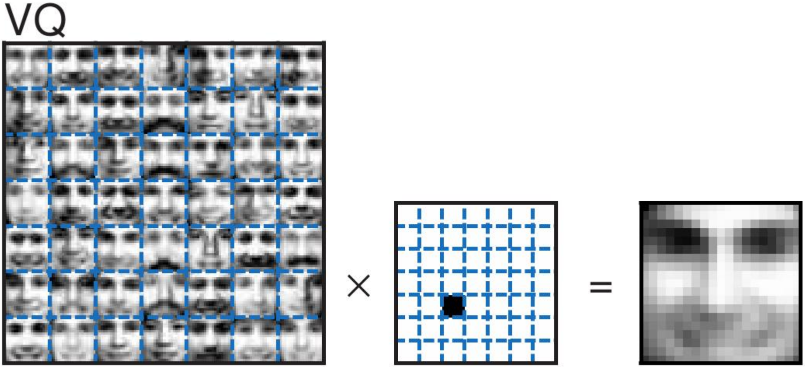
1. Set $a_k = \sum_i W_{ik}$
2. Divide $W_{ik}$ by $a_k$ for all $i$
3. Multiply $H_{kj}$ by $a_k$ for all $j$

Notice that this is does not change the matrix multiplication $WH$.

Interpretation: The $k$th *column* of $W$ can be interpreted as the $k$th *topic*. The $j$th *column* of $H$ can be interpreted as how much document $j$ uses each topic.

# NMF and Face Modeling

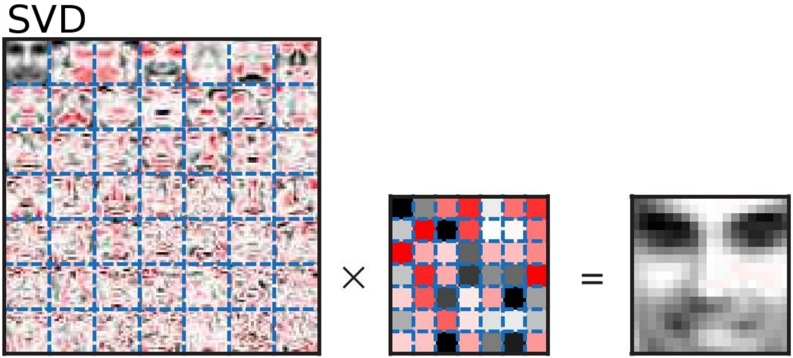
For face modeling, put the face images along the columns of $X$ and factorize. Show columns of $W$ as image. Compare this with K-means and SVD.



K-means (i.e., VQ): Equivalent to each column of $H$ having a single 1.
K-means learns averages of full faces.

# NMF and Face Modeling

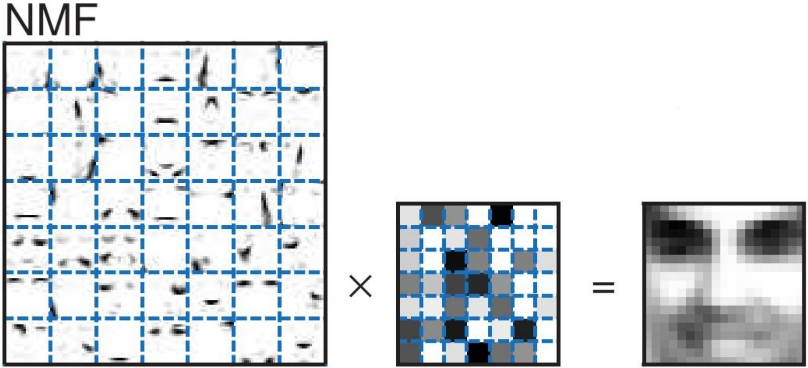
For face modeling, put the face images along the columns of $X$ and factorize. Show columns of $W$ as image. Compare this with K-means and SVD.



SVD: Finds the singular value decomposition of $X$.
Results not interpretable because of ± values and orthogonality constraint

# NMF and Face Modeling

For face modeling, put the face images along the columns of $X$ and factorize. Show columns of $W$ as image. Compare this with K-means and SVD.



NMF learns a “parts-based” representation. Each column captures something interpretable. This is a result of the nonnegativity constraint.