Bayesian Learning and Variational Inference

INTELLIGENT SYSTEMS FOR PATTERN RECOGNITION (ISPR)

DAVIDE BACCIU – DIPARTIMENTO DI INFORMATICA - UNIVERSITA’ DI PISA

DAVIDE.BACCIU@UNIPI.IT

Outline and Motivations

- Introduce the basic concepts of variational learning useful for both generative models and deep learning
- Bayesian latent variable models
  - A class of generative models for which variational or approximated methods are needed
- Latent Dirichlet Allocation
  - Possibly the simplest Bayesian latent variable model
  - Many applications in unsupervised text analytics, machine vision, ...
- A very quick intro to variational EM

Disclaimer

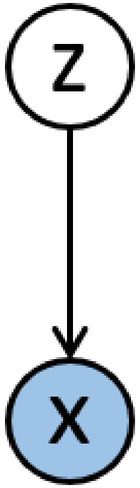
This lecture contains quite some

$$\int_S \dots, \sum_k \dots, \Gamma() \dots, \mathbb{E}[\log P] \dots$$

..try not to end up like this:

Problem Setup

Latent Variable Models



Latent variables

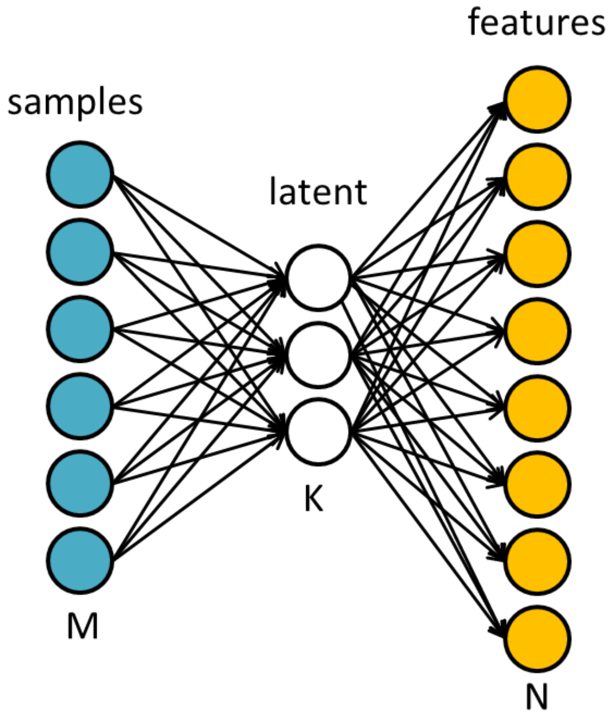
- Unobserved RV that define a hidden generative process of observed data
- Explain complex relation between many observable variables
- E.g. hidden states in HMM/CRF

Latent variable models likelihood

$$P(x) = \int_{\mathbf{z}} \prod_{i=1}^{N} P(x_i|\mathbf{z})P(\mathbf{z})d\mathbf{z}$$

Latent Space

Define a latent space where high-dimensional data can be represented



Assumption

Latent variables conditional and marginal distributions are more tractable than the joint distribution $P(\mathcal{X})$ (e.g $K \ll N$)

Tractability

- Introducing hidden variables can produce couplings between the distributions (i.e., one depending on the other) which can make their posterior intractable
- Bayesian learning introduces priors which introduce integrals in the posterior computations which are not always analytically or computationally tractable

This lecture is about how we can approximate such intractable problems
- Variational view of EM (also used in variational DL)

Kullback-Leibler (KL) Divergence

An information theoretic measure of closeness of two distributions p and q

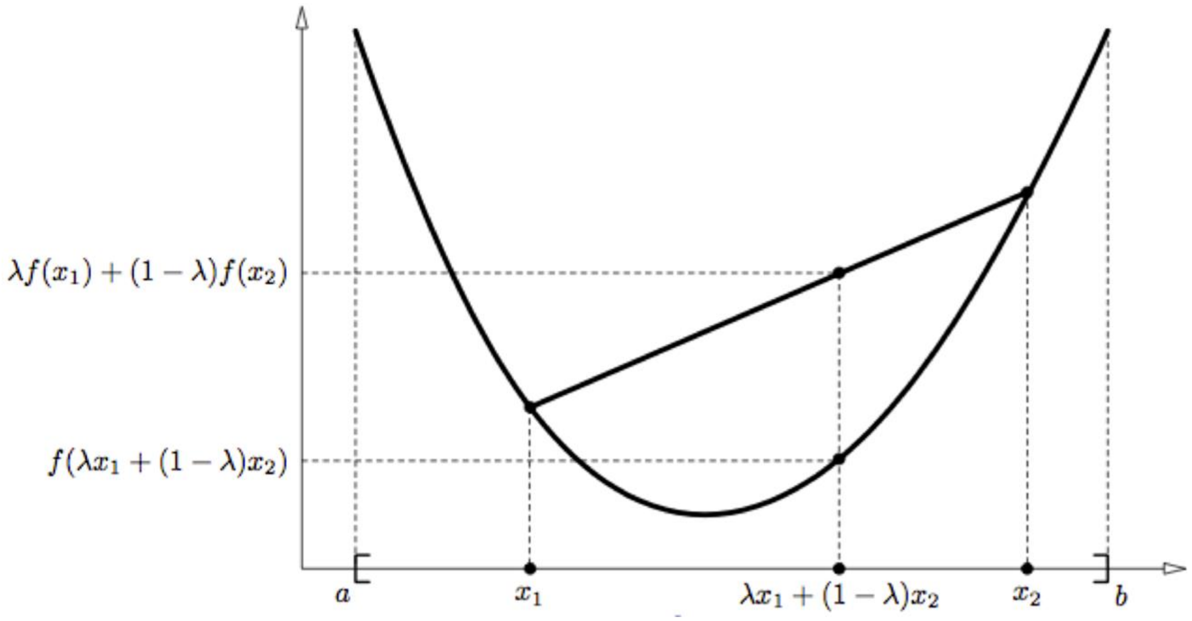
$$KL(q||p) = \mathbb{E}_q \left[\log \frac{q(z)}{p(z|x)} \right] = \langle \log q(z) \rangle_q - \langle \log p(z|x) \rangle_q$$

Note:

- A specialized definition for our latent variable setting
  - If q high and p high $\Rightarrow$ happy
  - If q high and p low $\Rightarrow$ unhappy
  - If q low $\Rightarrow$ don’t care (due to expectation)
- Its a divergence $\Rightarrow$ it is not symmetric

Jensen Inequality

Property of linear operators on convex/concave functions



Generalizes to

$$\frac{\sum_i a_i f(x_i)}{\sum_i a_i} \geq f\left(\frac{\sum_i a_i x_i}{\sum_i a_i}\right)$$

Applied in our probabilistic setting

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$$

since $f = log$ is concave

$$\lambda f(x) + (1-\lambda)f(x) \geq f(\lambda x + (1-\lambda)x)$$

Bounding Log-Likelihood with Jensen

The log-likelihood for a model with a single hidden variable Z and parameters θ (assume single sample for simplicity) is

$$\log P(x|\theta) = \log \int_z P(x,z|\theta)dz = \log \int_z \frac{Q(z|\phi)}{Q(z|\phi)} P(x,z|\theta)dz$$

which holds for $Q(z|\phi) \neq 0$ with parameters ϕ

Given the definition of expectation this rewrites as (Jensen)

$$\log P(x|\theta) = \log \mathbb{E}_Q\left[\frac{P(x,z)}{Q(z)}\right] \geq \mathbb{E}_Q\left[\log \frac{P(x,z)}{Q(z)}\right]$$

$$= \underbrace{\mathbb{E}_Q[\log P(x,z)]}_{\text{Expectation of Joint Distribution}} - \underbrace{\mathbb{E}_Q[\log Q(z)]}_{\text{Entropy}} = \mathcal{L}(x,\theta,\phi)$$

How Good is this Lower Bound?

$$\log P(x|\theta) - \mathcal{L}(x,\theta,\phi) = ?$$

Inserting the definition of $\mathcal{L}(x,\theta,\phi)$

$$\log P(x|\theta) - \int_z Q(z) \log \frac{P(x,z)}{Q(z)} dz$$

Introducing $Q(z)$ by marginalization ($\int_z Q(z) = 1$)

$$\int_z Q(z) \log P(x) dz - \int_z Q(z) \log \frac{P(x,z)}{Q(z)} dz =$$

$$\int_z Q(z) \log \frac{P(x)Q(z)}{P(x,z)} dz$$

How Good is this Lower Bound?

$$\log P(x|\theta) - \mathcal{L}(x, \theta, \phi) = ?$$

Inserting the definition of $\mathcal{L}(x, \theta, \phi)$

$$\log P(x|\theta) - \int_z Q(z) \log \frac{P(x,z)}{Q(z)} dz$$

Introducing $Q(z)$ by marginalization ($\int_z Q(z) = 1$)

$$\int_z Q(z) \log P(x) dz - \int_z Q(z) \log \frac{P(x,z)}{Q(z)} dz =$$

$$\mathbb{E}_Q \left[\log \frac{Q(z)}{P(z|x)} \right] \quad = KL\big(Q(z|\phi) || P(z|x,\theta)\big)$$

Defining and Interpreting the Bound

We can assume the existence of a probability $Q(z|\phi)$ which allows to bound the likelihood $P(x|\theta)$ from below using $\mathcal{L}(x, \theta, \phi)$

The term $\mathcal{L}(x, \theta, \phi)$ is called **variational bound** or **evidence lower bound (ELBO)**

The optimal bound is obtained for $KL\big(Q(z|\phi)||P(z|x, \theta)\big) = 0$, that is if we choose $\boldsymbol{Q(z|\phi) = P(z|x, \theta)}$

Minimizing KL is equivalent to maximize the ELBO ⇒ change a sampling problem with an optimization problem

Variational View of Expectation Maximization

EM Learning Reformulated

Maximum likelihood learning with hidden variables can be approached by maximization of the ELBO

$$\max_{\theta,\phi} \sum_{n=1}^{N} \mathcal{L}(x_n, \theta, \phi)$$

where θ are the model parameters and ϕ serve in $Q(z|\phi)$

- If $P(z|x,\theta)$ is tractable $\Rightarrow$ use it as $Q(z|\phi)$ (optimal ELBO)
- O.w. choose $Q(z|\phi)$ as a tractable family of distributions
  - find ϕ that minimize $KL(Q(z|\phi)||P(z|x,\theta))$, or
  - find ϕ that maximize $\mathcal{L}(\cdot,\phi)$

A Generative Model for Multinomial Data

A **Bag of Words** (BOW) representation of a document is the classical example of multinomial data (for text, images, graphs,...)

A BOW dataset (corpora) is the $N \times M$ term-document matrix

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1i} & \cdots & x_{1M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{j1} & \cdots & x_{ji} & \cdots & x_{jM} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{N1} & \cdots & x_{Ni} & \cdots & x_{NM} \end{bmatrix}$$

- N : number of vocabulary items w_j
- M : number of documents d_i
- $x_{ij} = n(w_j, d_i)$: number of occurrences of w_j in d_i

IN SUPREMÆ DIGNITATIS 1343 UNIVERSITÀ DI PISA

Documents as Mixtures of Latent Variables

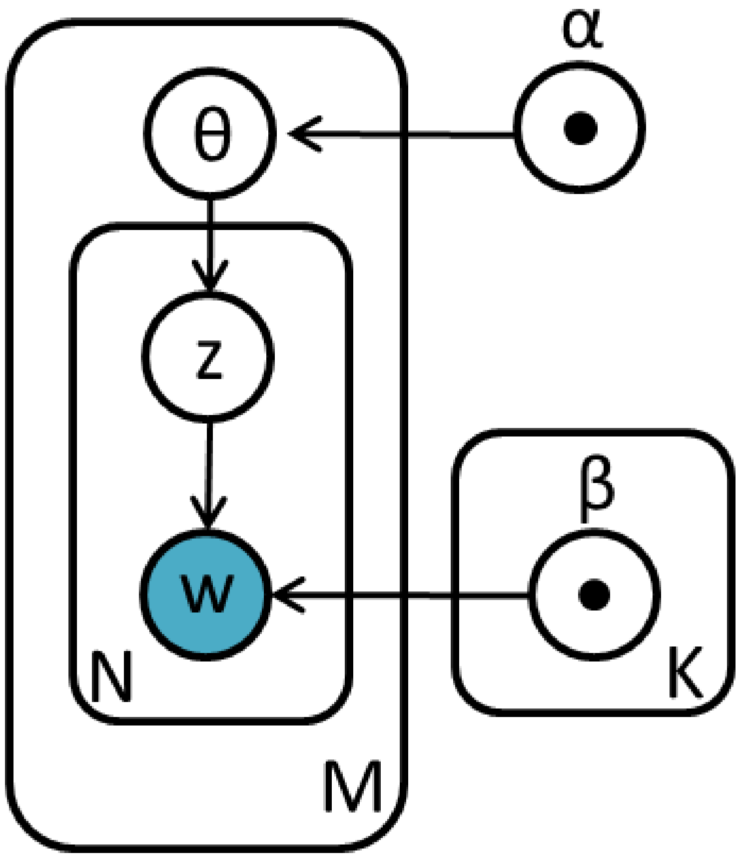
Latent topic models consider documents (i.e. item containers) as a mixture of topics

- A topic identifies a pattern in the co-occurrence of multinomial items w_j within the documents
- Mixture of topics ⇒ Associate an interpretation (topic) to each item in a document, whose interpretation is then a mixture of the items' topics

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1i} & \cdots & x_{1M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{j1} & \cdots & x_{ji} & \cdots & x_{jM} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{N1} & \cdots & x_{Ni} & \cdots & x_{NM} \end{bmatrix}$$

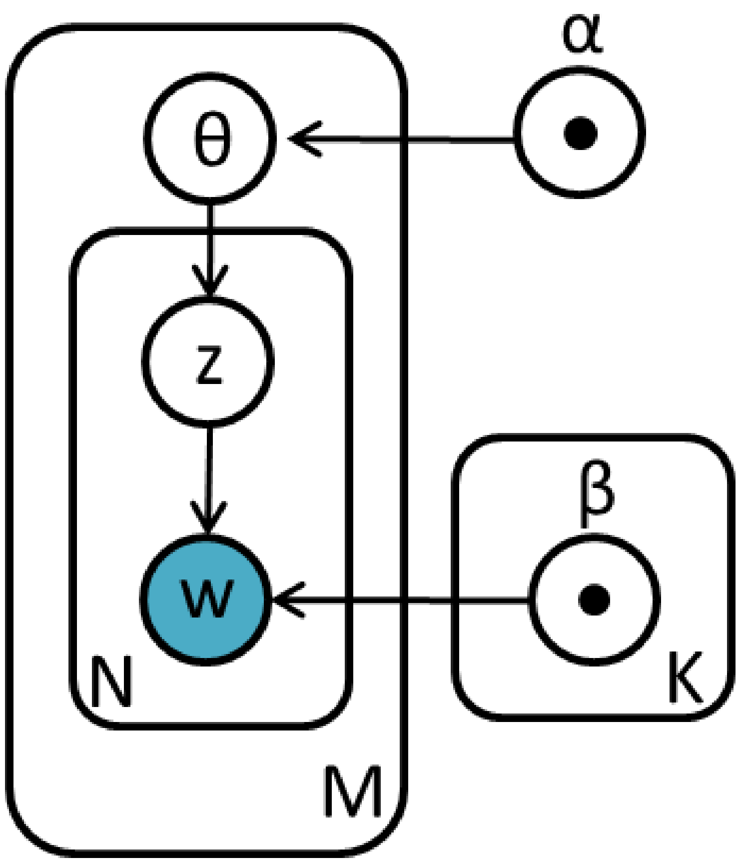
UNIVERSITÀ DI PISA

Latent Dirichlet Allocation (LDA)



- LDA models a document as a mixture of topics z
  - Assigning one topic z to each item w with probability $P(w|z,\beta)$
  - Pick one topic for the the whole document with probability $P(z|\theta)$
- Key point - Each document has its personal topic proportion θ sampled from a distribution
  - θ defines a multinomial distribution but it is a random variable as well

LDA Distributions



- $P(w|z,\beta)$ multinomial item-topic distribution
- $P(z|\theta)$ multinomial topic distribution with document-specific parameter θ
- $P(\theta|\alpha)$ Dirichlet distribution with hyperparameter α
  - A distribution for vectors that sum to 1 (simplex)
  - The elements of a multinomial are vector that sum to 1!

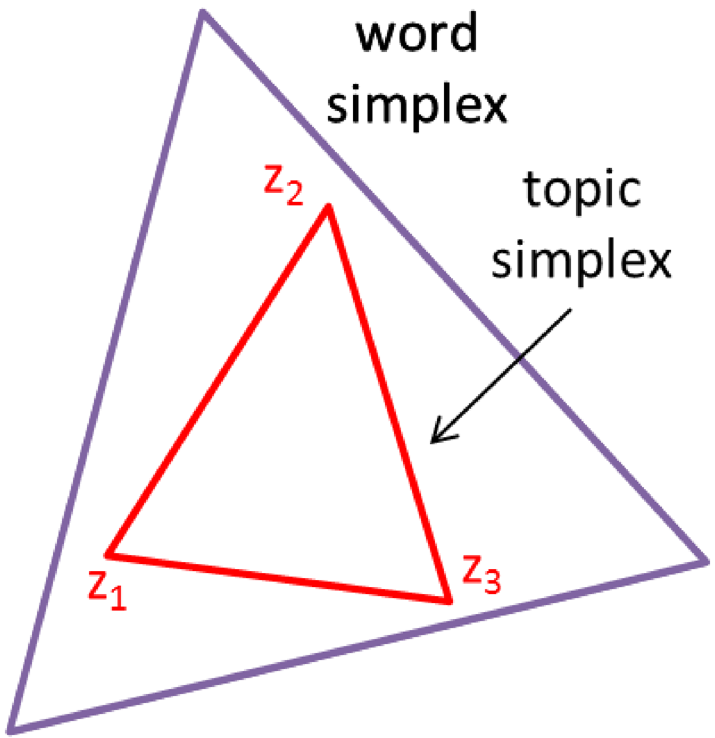
Dirichlet Distribution

- Why a Dirichlet distribution?
  - Conjugate prior to multinomial distribution
  - If the likelihood is multinomial with a Dirichlet prior then posterior is Dirichlet
- Dirichlet distribution

$$P(\theta|\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

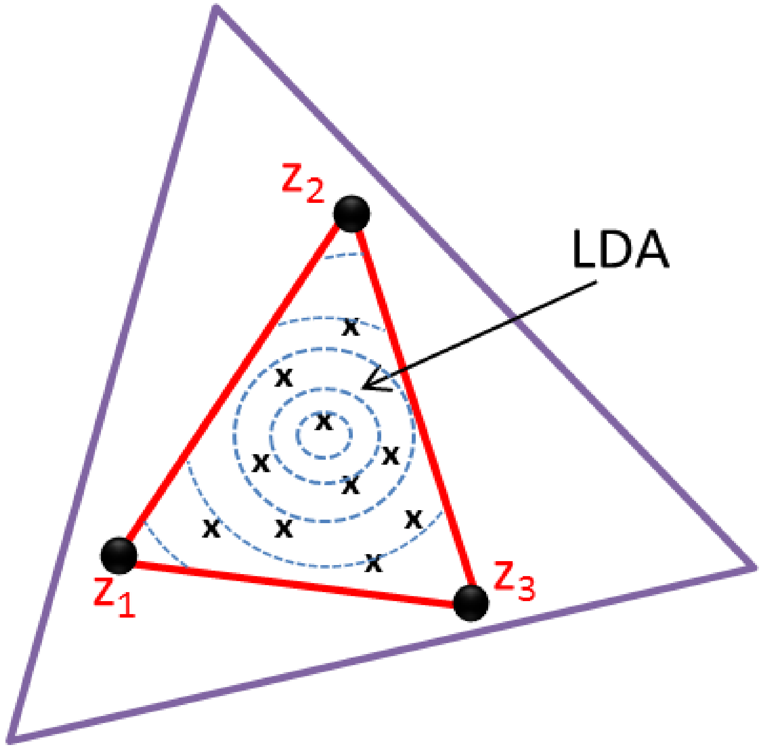
  - Dirichlet parameter α_k is a prior count of the k-th topic
  - It controls the mean shape and sparsity of multinomial parameters θ

Geometric Interpretation



LDA finds a set of K projection functions on the K-dimensional topic simplex

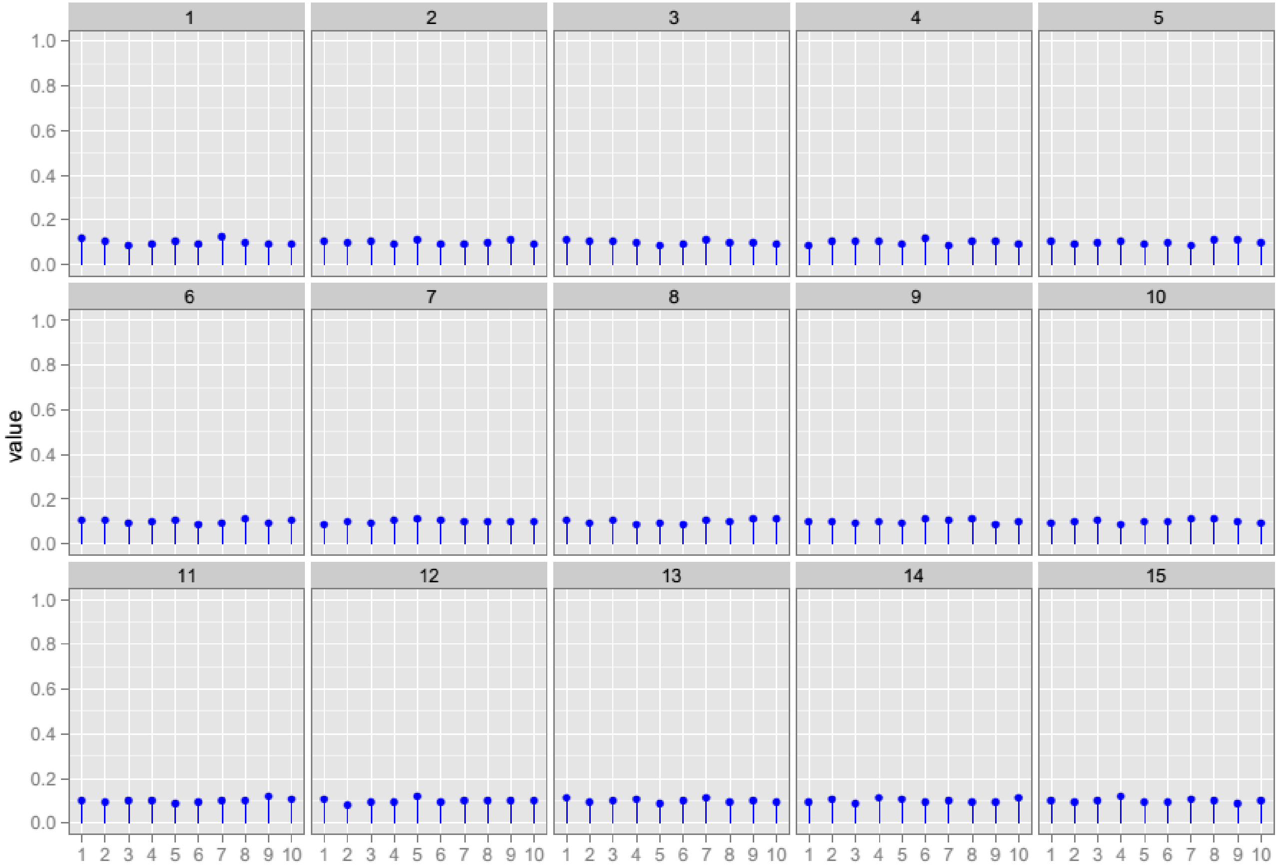
Geometric Interpretation



LDA finds a set of K projection functions on the K-dimensional topic simplex

Effect of the α parameter

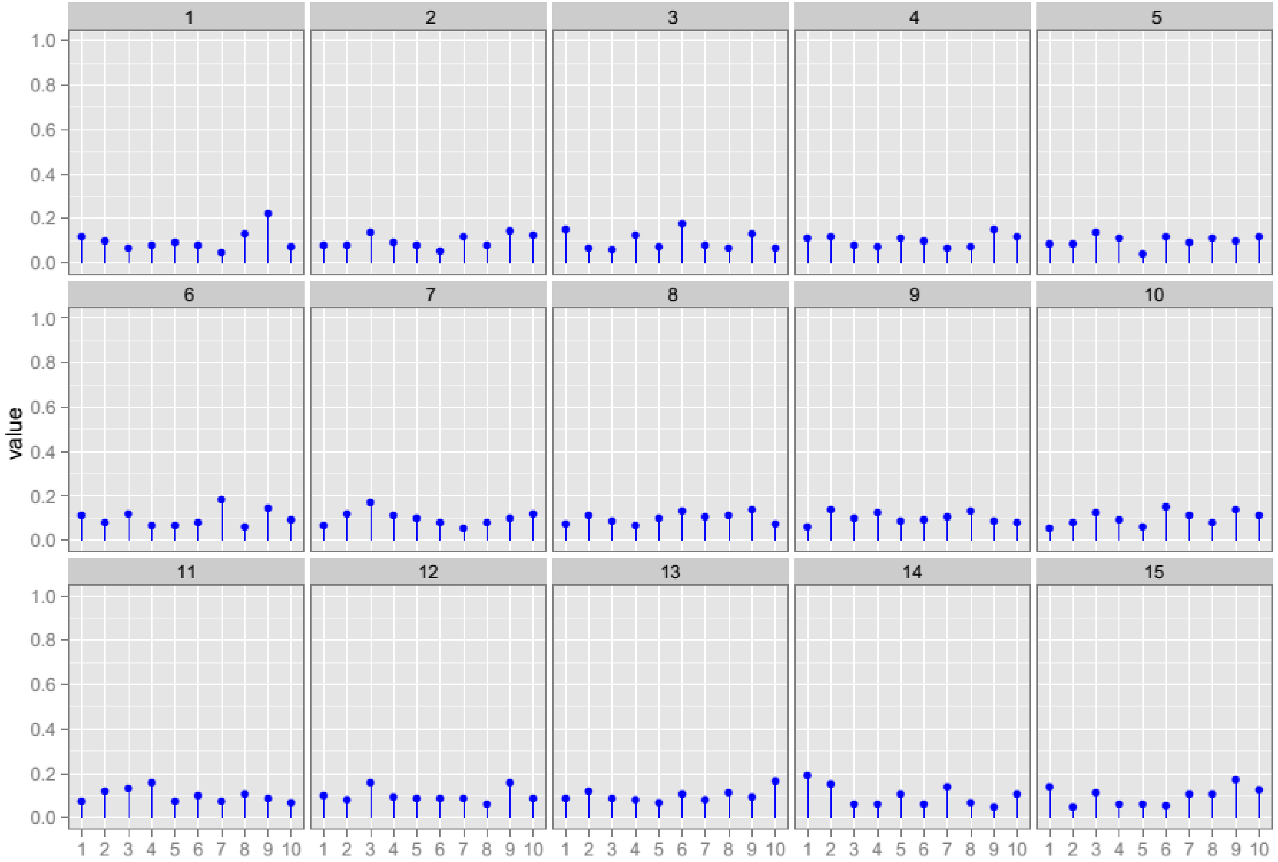
$\alpha = 100$



Slide Credit - Blei at KDD 2011 Tutorial

Effect of the α parameter

$\alpha = 10$

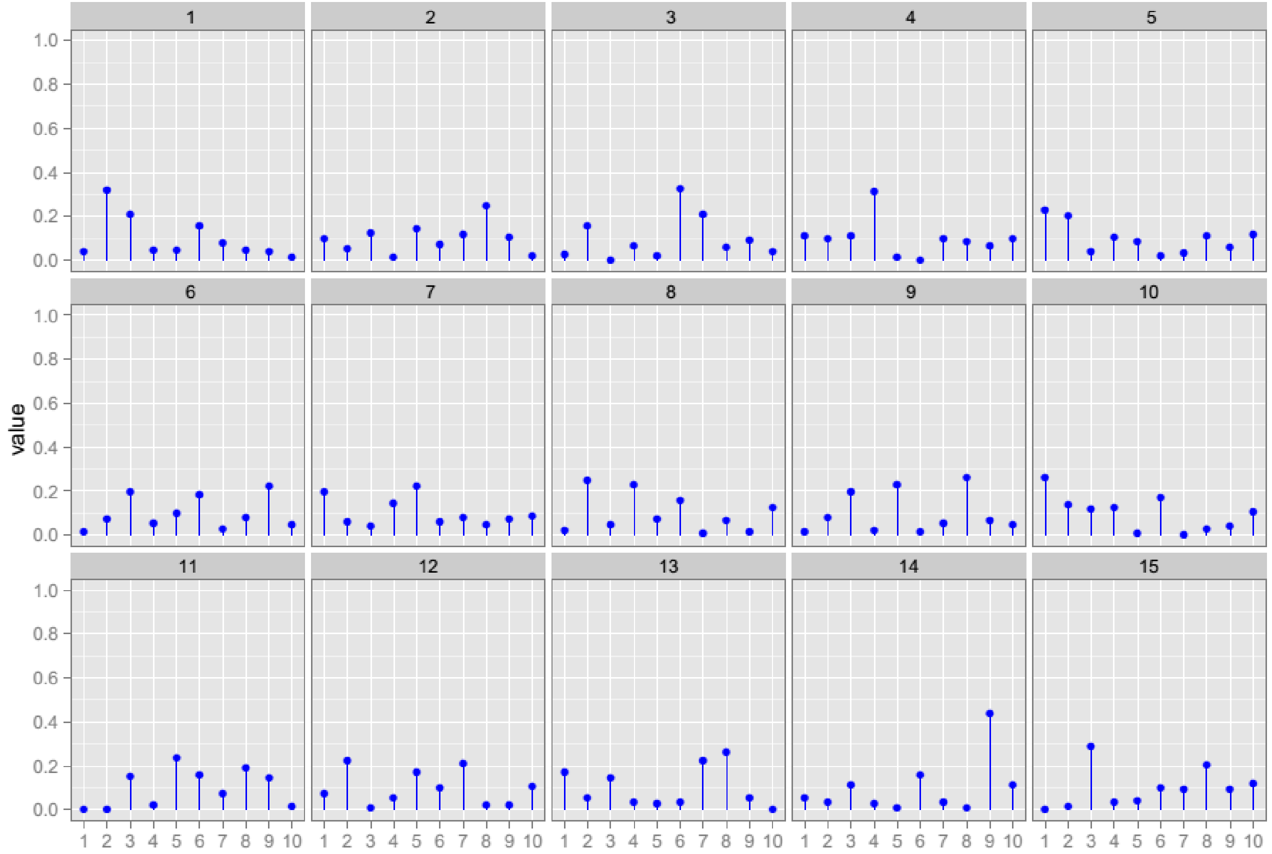


Slide Credit - Blei at KDD 2011 Tutorial

UNIVERSITÀ DI PISA

Effect of the α parameter

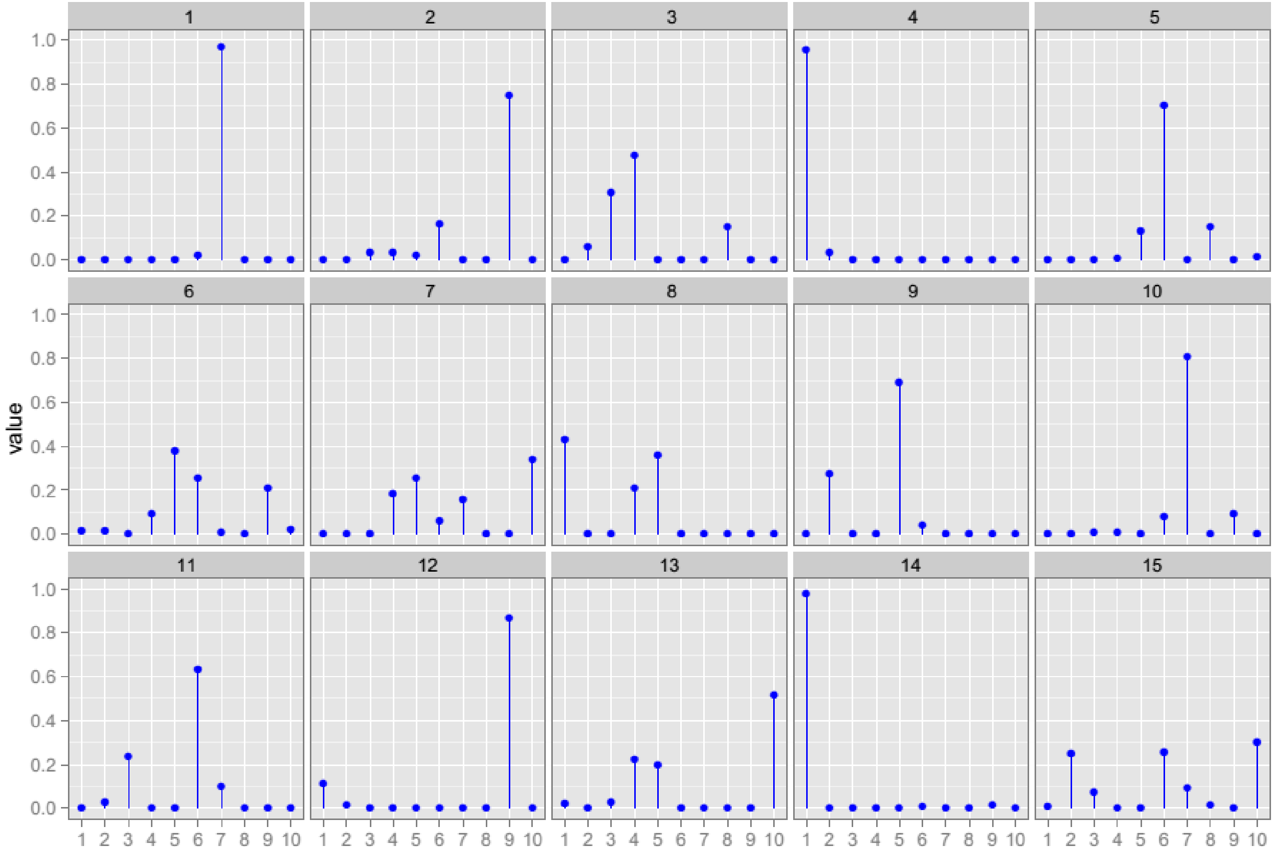
$\alpha = 1$



Slide Credit - Blei at KDD 2011 Tutorial

Effect of the α parameter

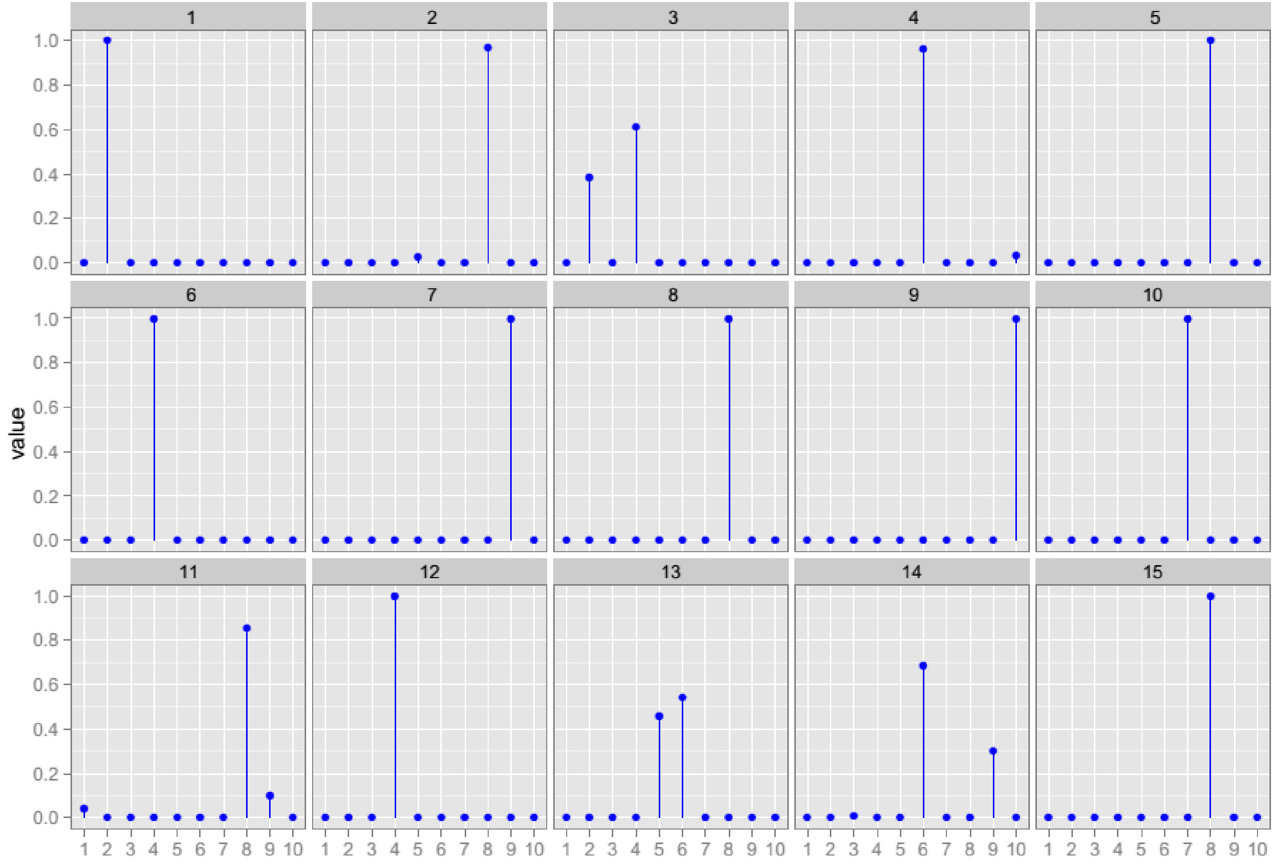
$\alpha = 0.1$



Slide Credit - Blei at KDD 2011 Tutorial

Effect of the α parameter

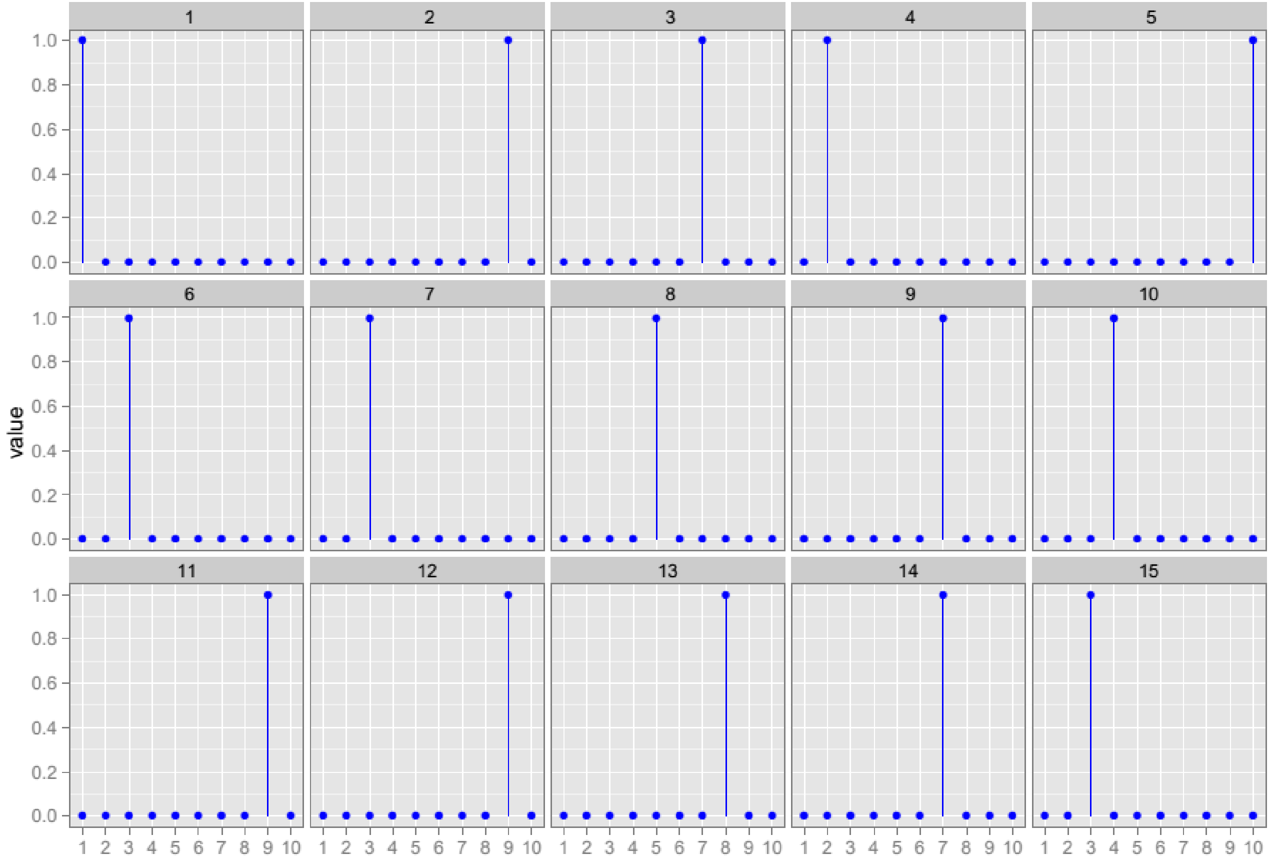
$\alpha = 0.01$



Slide Credit - Blei at KDD 2011 Tutorial

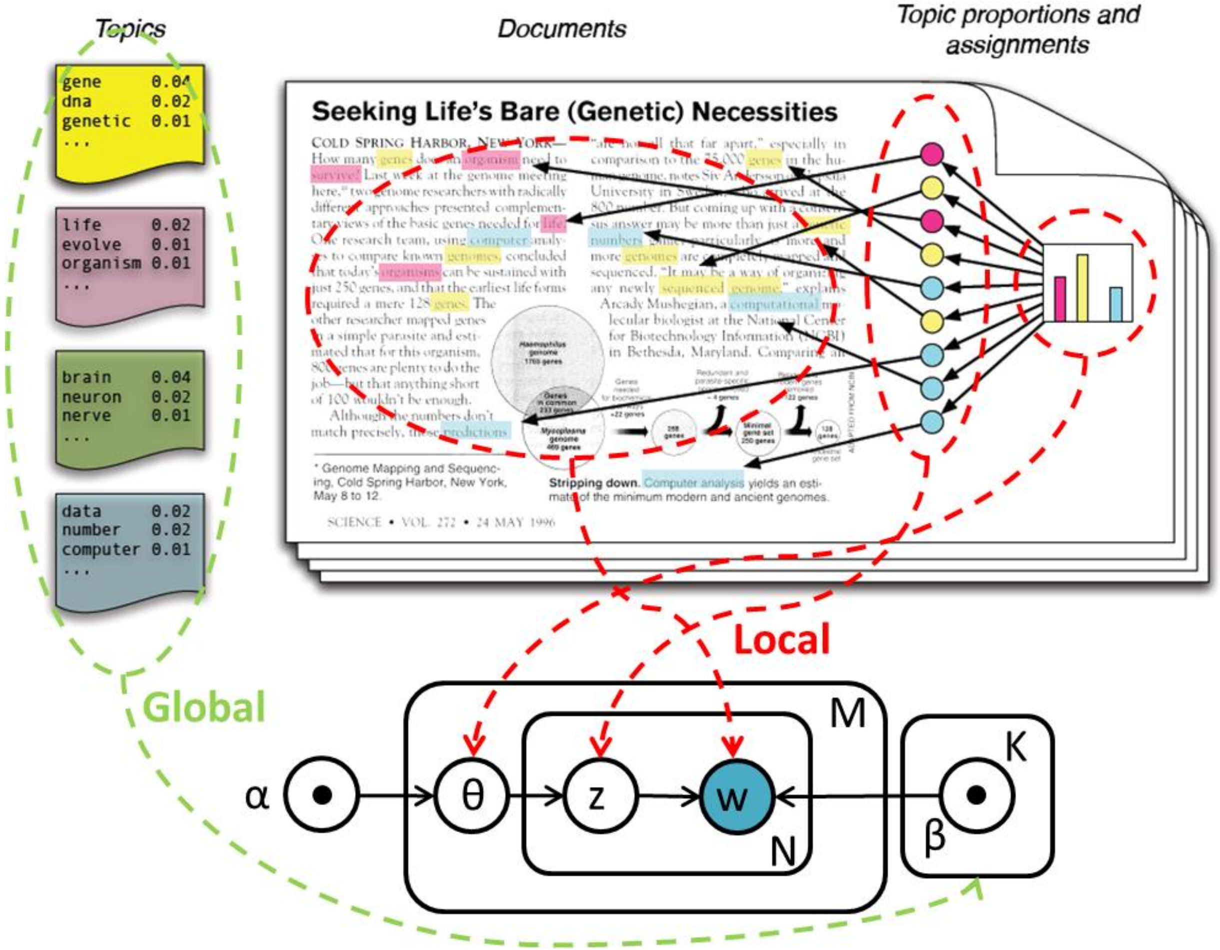
Effect of the α parameter

$\alpha = 0.001$

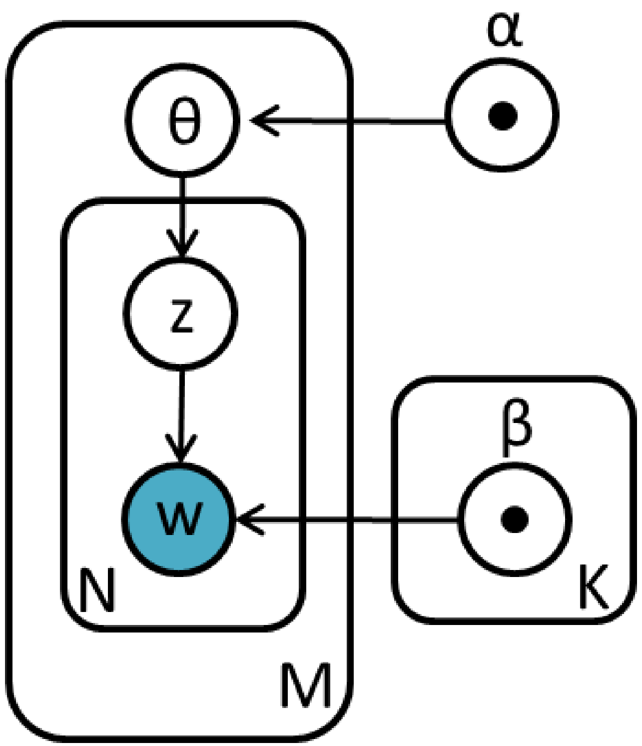


Slide Credit - Blei at KDD 2011 Tutorial

LDA and Text Analysis



LDA Generative Process



For each of the M documents

- Choose $\theta \sim$ Dirichlet(α)
- For each of the N items
  - Choose a topic $z \sim$ Multinomial(θ)
  - Pick an item w_j with multinomial probability $P(w_j|z,\beta)$

Multinomial topic-item parameter matrix $[\beta]_{K \times V}$

$$\beta_{kj} = P(w_j = 1|z_k = 1) \text{ or } P(w_j = 1|z = k)$$

$$P(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = P(\theta|\alpha) \prod_{j=1}^{N} P(z_j|\theta) P(w_j\,|z_j\,, \beta)$$

Learning in LDA

Marginal distribution (a.k.a. likelihood) of a document $d = \boldsymbol{w}$

$$P(\boldsymbol{w}|\alpha,\beta) = \int \sum_{\boldsymbol{z}} P(\theta,\boldsymbol{z},\boldsymbol{w}|\alpha,\beta)d\theta = \int P(\theta|\alpha) \prod_{j=1}^{N} \sum_{z_j=1}^{k} P(z_j|\theta)P(w_j|z_j,\beta)d\theta$$

Given $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M\}$, find (α,β) maximizing

$$\mathcal{L}(\alpha,\beta) = \log \prod_{i=1}^{M} P(\boldsymbol{w}_i|\alpha,\beta)$$

Learning with hidden variables $\Longrightarrow$ Expectation-Maximization

Key problem is inferring latent variables posterior

$$P(\theta,\boldsymbol{z}|\boldsymbol{w},\alpha,\beta) = \frac{P(\theta,\boldsymbol{z},\boldsymbol{w}|\alpha,\beta)}{P(\boldsymbol{w}|\alpha,\beta)}$$

Posterior Inference

- Optimal ELBO is achieved when Q(z) is equal to the latent variable posterior

$$P(\theta, \boldsymbol{z} | \boldsymbol{w}, \alpha, \beta) = \frac{P(\theta, \boldsymbol{z}, \boldsymbol{w} | \alpha, \beta)}{P(\boldsymbol{w} | \alpha, \beta)}$$

- Key problem is that computation of the posterior is not tractable
- Computation of the denominator is intractable due to the couplings between β and θ in the summation over topics

$$P(\boldsymbol{w} | \alpha, \beta) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \int \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \left(\prod_{j=1}^{N} \sum_{k=1}^{K} \prod_{v=1}^{V} (\theta_k \beta_{kv})^{w_j^v} \right) d\theta$$

Approximating Parameter Inference in LDA

Variational Inference

- Maximize the variational bound without using the optimal posterior solution
  - Write a $Q(\mathbf{z}|\phi)$ function that is sufficiently similar to the posterior but tractable
  - $Q(\mathbf{z}|\phi)$ should be such that β and θ are no longer coupled
  - Fit ϕ parameter so that $Q(\mathbf{z}|\phi)$ is close to $P(\mathbf{w}|\alpha, \beta)$ according to KL
- Variational LDA: Blei, Ng and Jordan, 2003
- Fast convergence (but it is an approximation)

Sampling Approach

- Construct a Markov chain on the hidden variables whose limiting distribution is the posterior
- Sampling LDA: Griffiths and Steyvers, 2004
- Slow convergence (but it is as accurate as you whish)

UNIVERSITÀ DI PISA

Variational Inference

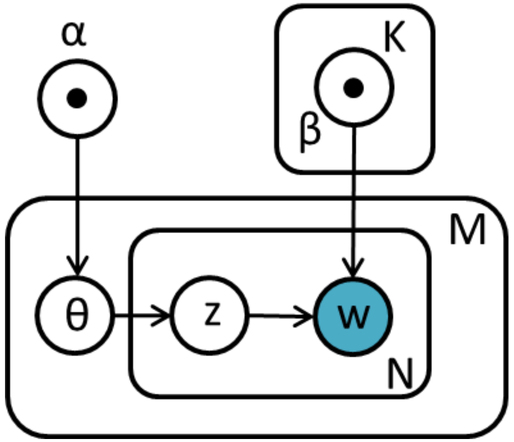
Key Idea

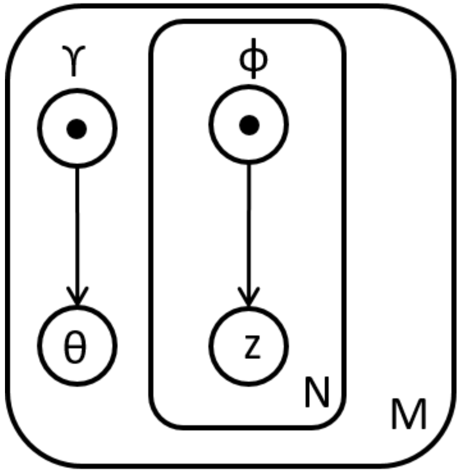
Assume that our distribution $Q(z|\phi)$ factorizes (it is tractable) → mean-field assumption

$$Q(\mathbf{z}|\phi) = Q(z_1, \ldots, z_K|\phi) = \prod_{k=1}^{K} Q(z_k \,|\phi_k)$$

- Can be made more general by factorizing on groups of latent variables
- Does not contain the true posterior because hidden variables are dependent
- Variational inference
  - Optimize ELBO using $Q(\mathbf{z}|\phi)$ factorized distribution
  - Coordinate ascent inference - Iteratively optimize each variational distribution holding the others fixed

Variational LDA Distribution





Given $\Phi = \{\gamma, \phi\}$ as variational approximation parameters

$$Q(\theta, \mathbf{z}|\Phi) = Q(\theta|\gamma) \prod_{n=1}^{N} Q(z_n|\phi_n)$$

Then we have the model parameters β of sample distribution $P(\theta, \mathbf{z}, \mathbf{w}|\ \beta)$

Variational Expectation-Maximization

Find the Φ, Ψ that maximize the ELBO

$$\mathcal{L}(\boldsymbol{w}, \Phi, \beta) = \mathbb{E}_Q[\log P(\theta, \boldsymbol{z}, \boldsymbol{w}|\beta)] - \mathbb{E}_Q[\log Q(\theta, \boldsymbol{z}|\Phi)]$$

by alternate maximization

1. repeat
2. Fix β: update variational parameters Φ^* (E-STEP)
3. Fix $\Phi = \Phi^*$: update model parameters β^* (M-STEP)
4. until little likelihood improvement

Unlike EM, variational EM has no guarantee to reach a local maximizer of $\mathcal{L}$

A (scary?) hidden slide

Expand the lower bound using P and Q factorization

$$L(\gamma,\phi;\alpha,\beta) = \mathrm{E}_q[\log p(\theta\,|\,\alpha)] + \mathrm{E}_q[\log p(\mathbf{z}\,|\,\theta)] + \mathrm{E}_q[\log p(\mathbf{w}\,|\,\mathbf{z},\beta)] - \mathrm{E}_q[\log q(\theta)] - \mathrm{E}_q[\log q(\mathbf{z})].$$

$$\begin{aligned} L(\gamma,\phi;\alpha,\beta) = {} & \log\Gamma\left(\textstyle\sum_{j=1}^k \alpha_j\right) - \sum_{i=1}^k \log\Gamma(\alpha_i) + \sum_{i=1}^k (\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^k \gamma_j\right)\right) \\ & + \sum_{n=1}^N \sum_{i=1}^k \phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^k \gamma_j\right)\right) \\ & + \sum_{n=1}^N \sum_{i=1}^k \sum_{j=1}^V \phi_{ni} w_n^j \log\beta_{ij} \\ & - \log\Gamma\left(\textstyle\sum_{j=1}^k \gamma_j\right) + \sum_{i=1}^k \log\Gamma(\gamma_i) - \sum_{i=1}^k (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^k \gamma_j\right)\right) \\ & - \sum_{n=1}^N \sum_{i=1}^k \phi_{ni}\log\phi_{ni}, \end{aligned}$$

Plug in the respective distributions

$$\phi_{ni} \propto \beta_{iv}\exp\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^k \gamma_j\right)\right)$$

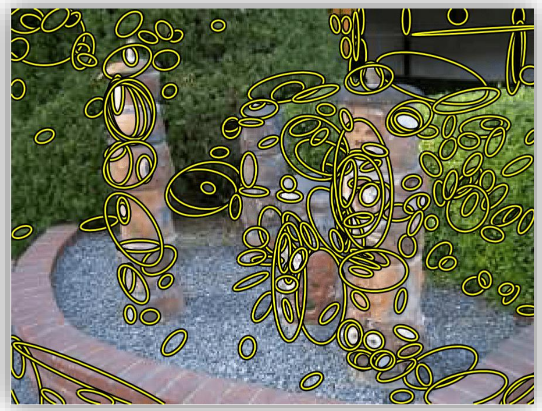
LDA Applications

- Why using latent topic models?
- Organize large collections of documents by identifying shared topics
- Understanding the documents semantics (unsupervised)
- Documents are of different nature
  - Text
  - Images
  - Video
  - Relational data (graphs, time-series, etc..)
- In short: a model for collections of high-dimensional vectors whose attributes are multinomial distributions

Understanding Image Collections

How can we apply latent topic analysis to visual documents?

- We need a way to represent visual content as in text
  - Text ≡ collection of discrete items ⇒ words
  - Image ≡ collection of discrete items ⇒ ?



- Visual patches
  - Feature detectors to identify relevant image parts (MSER)
  - Feature descriptors to represent content (SIFT)
  - How can I obtain a discrete vocabulary for visual terms?

Building a Visterm Vocabulary

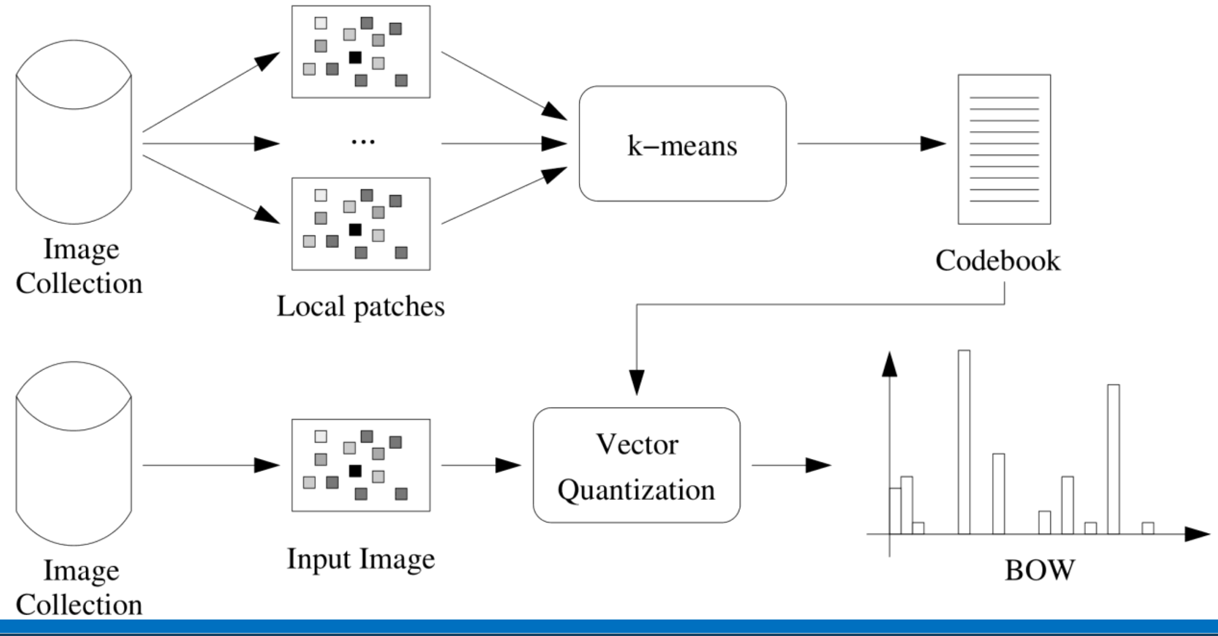
Given a dataset of images

1. For each image I
   - Identify interesting points (MSER/SIFT/grid)
   - Extract the corresponding descriptors (SIFT)
2. Concatenate the image descriptors in a $128 \times N$ matrix, where N is the total number of descriptors extracted
3. Cluster the descriptors in C groups to obtain a vocabulary of C visterms (k-means)

You know all the necessary techniques to build this system!

Representing Image as a Bag of Items

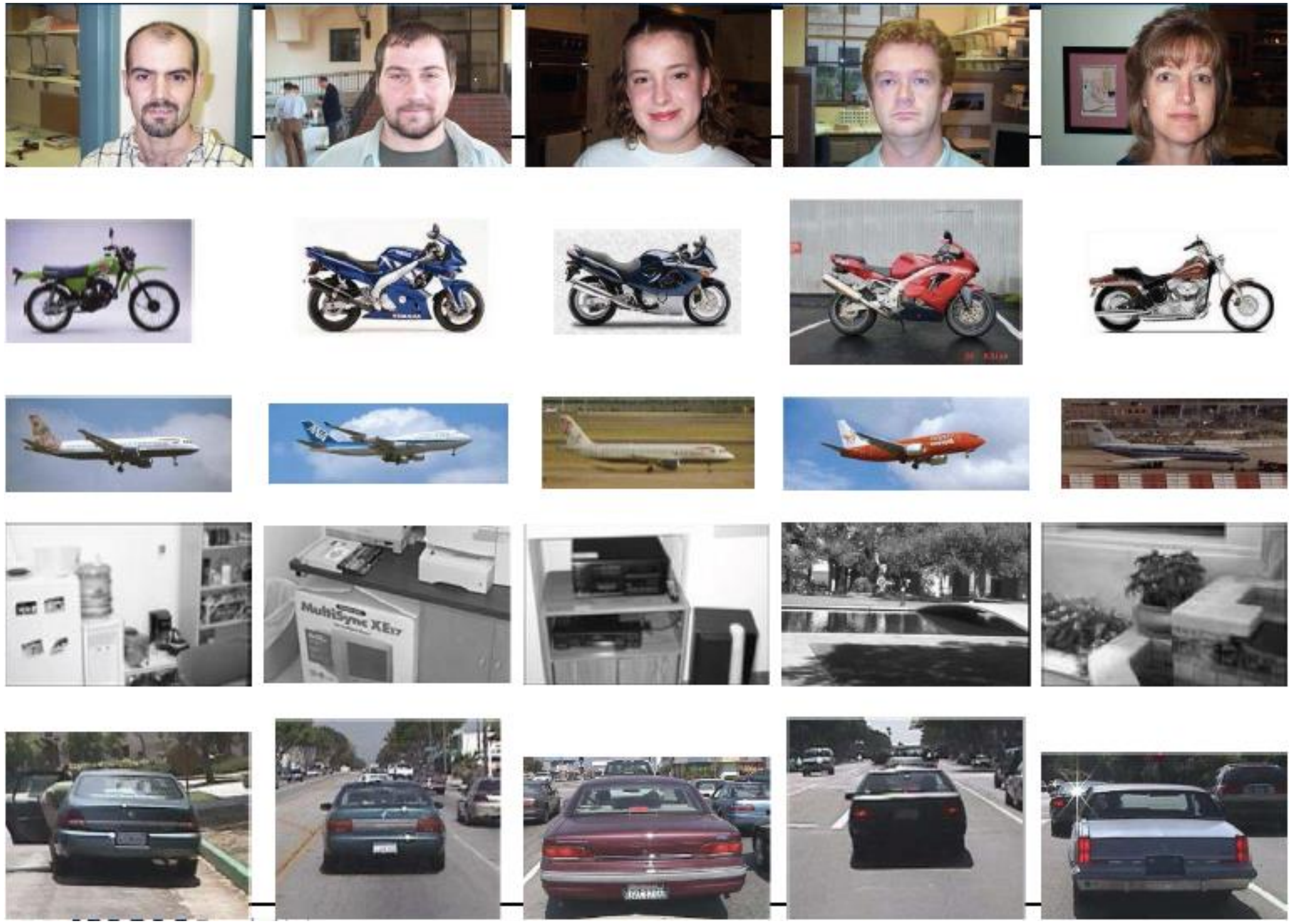
- Each image I is a document and each visual patch inside it is an item
- Associate each patch to the nearest cluster/visterm c
- Count the occurrences of each dictionary visterm c in your image
- Represent the image as a vector of visterm counts



LDA Image Understanding

Assigning a topic to each visual patch

LDA Image Understanding

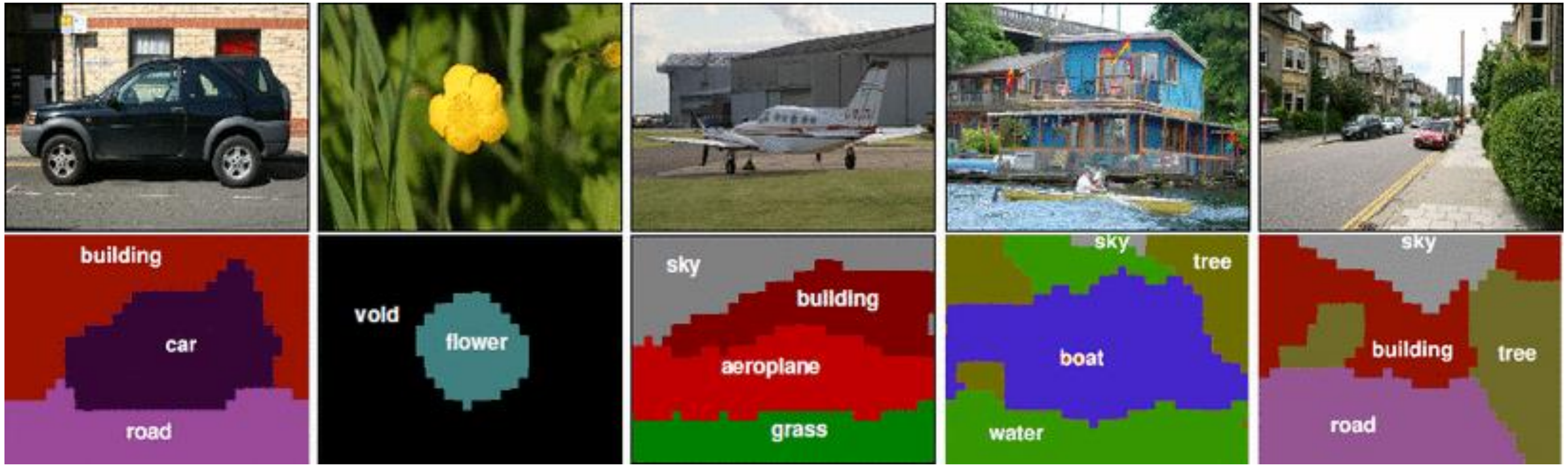


LDA Image Understanding

Unsupervised Semantic Segmentation

Combine latent topics with Markov random fields

- Use LDA to identify topics of some pixel patches
- Use MRF to diffuse LDA topics and enforce coherent pixel-level semantic segmentation

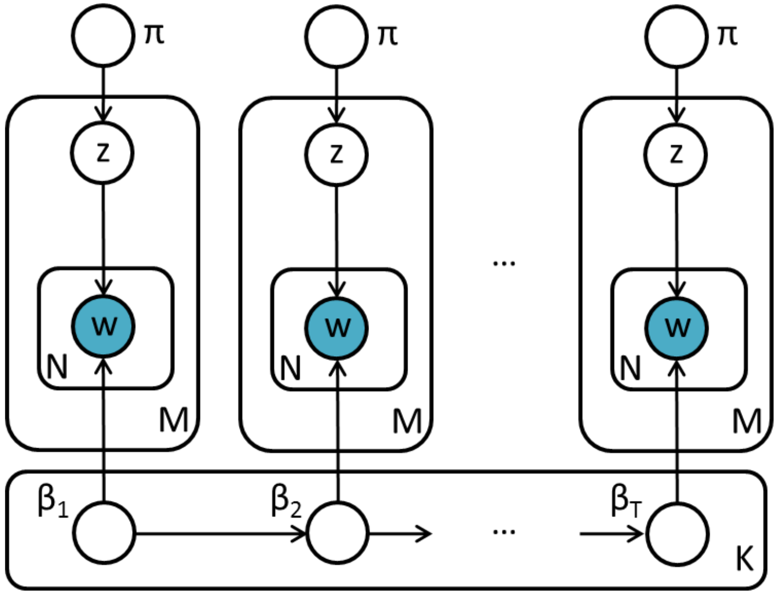


Zhao, Fei-Fei and Xing, Image Segmentation with Topic Random Field, ECCV 2010

Dynamical Topic Models

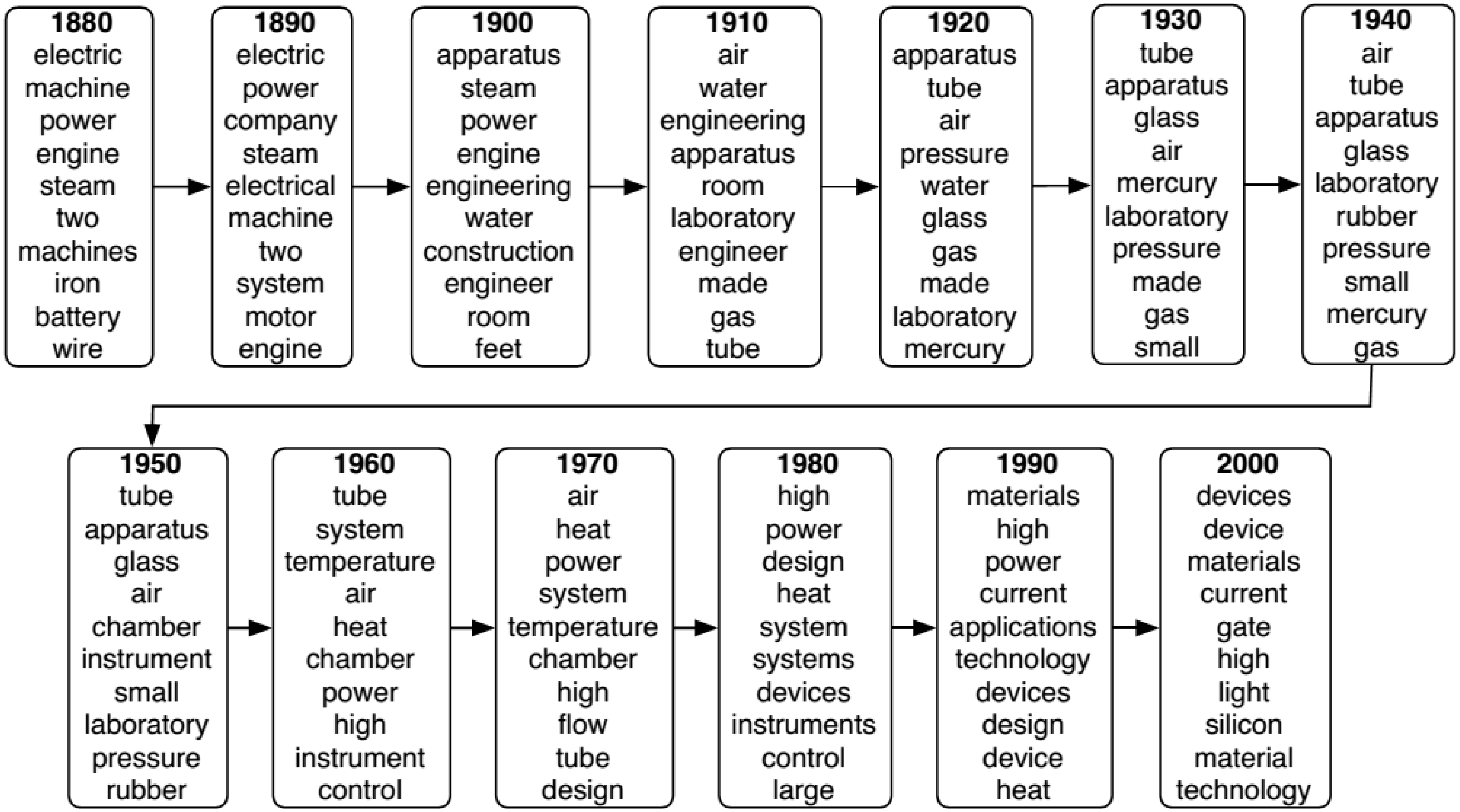
LDA assumes that the document order does not count

- What if we want to track topic evolution over time?
- Tracking how language changes over time
- Videos are image documents over time



Blei and Lafferty. Dynamic topic models, ICML 2006

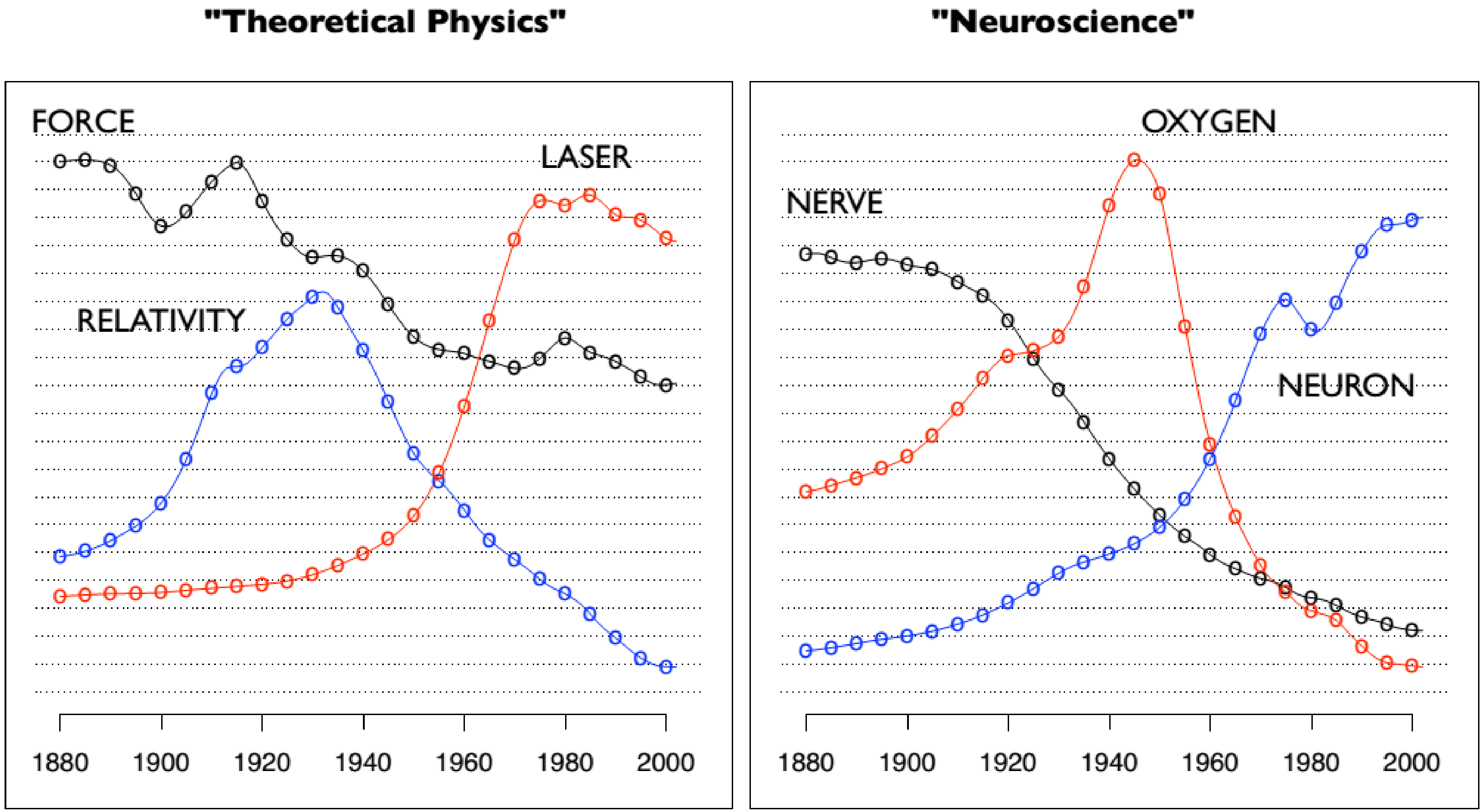
Topic Evolution over Time



https://github.com/blei-lab/dtm

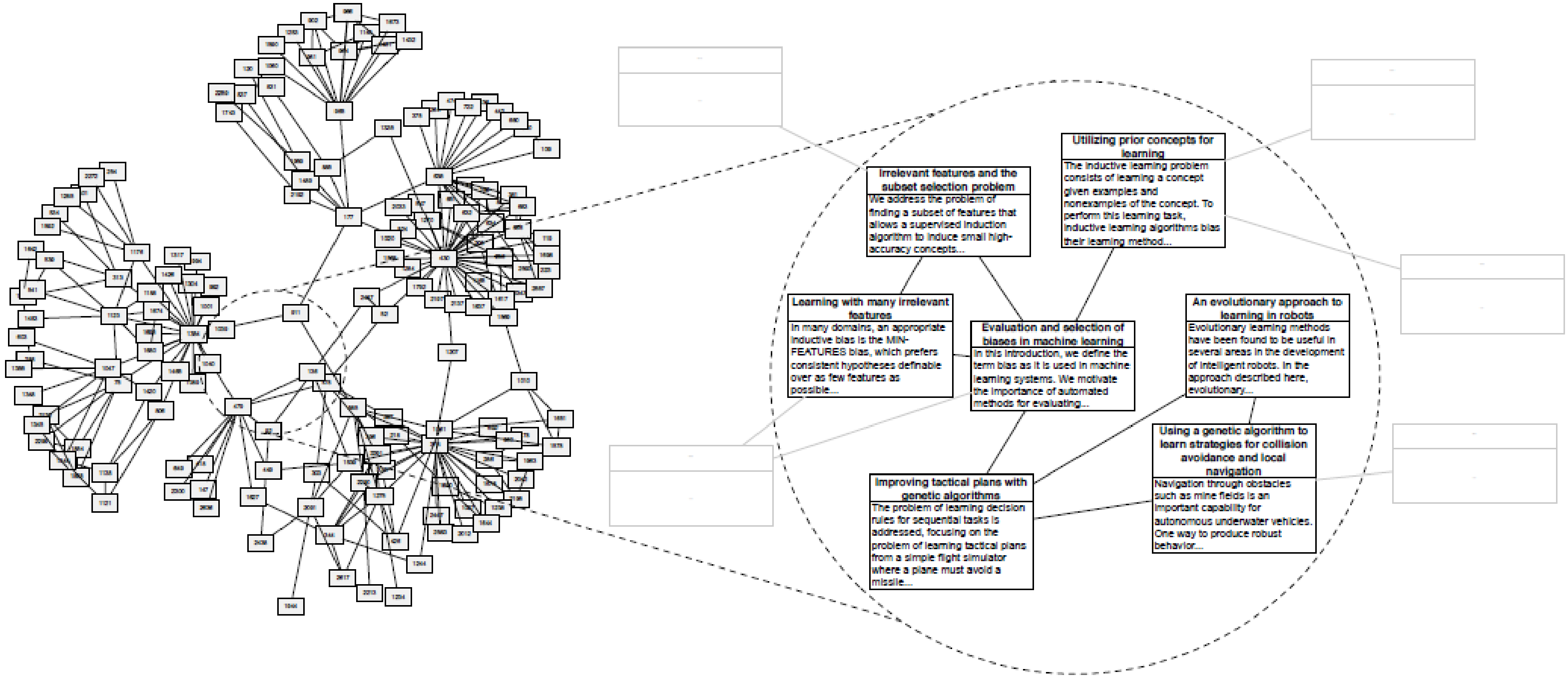
UNIVERSITÀ DI PISA

Topic Trends



https://github.com/blei-lab/dtm

Relational Topic Models



- Using topic models with relational data (graphs)
- Community discovery and connectivity pattern profiles (Kemp, Griffiths, Tenenbaum, 2004)
- Joint content-connectivity analysis (Blei,Chang, 2010)

Variational Learning in Code

- PyMC3 - Python library with particular focus on variational algorithms (not PyMC!)
- Edward - Python library with lots of variational inference from the father of LDA
- Bayespy - Variational Bayesian inference for conjugate-exponential family only
- Autograd - Variational and deep learning with differentiation as native Python operator (no strange backend)
- Matlab does not have official support for variational learning but standalone implementation of various models (check Variational-Bayes.org)
- LDA is implemented in many Python libraries: scikit-learn, pypi, gensim (efficient topic models)

Take Home Messages

- Bayesian learning amounts to treating distributions as random variables sampled from another distribution
  - Add priors to ML distributions
  - Learn functions instead of point estimates
- Latent Dirichlet Allocation
  - Bayesian model to organize collections of multinomial data
  - Unsupervised latent representation learning
- Variational lower bound
  - Maximizing a lower bound of an intractable likelihood
  - Alternatively estimate variational parameters and maximize w.r.t model parameters
  - A fundamental concept to understand variational deep learning

Next Lecture

Sampling Methods

- Introduction to sampling methods
- Ancestral sampling
- Gibbs Sampling
- MCMC family and advanced methods