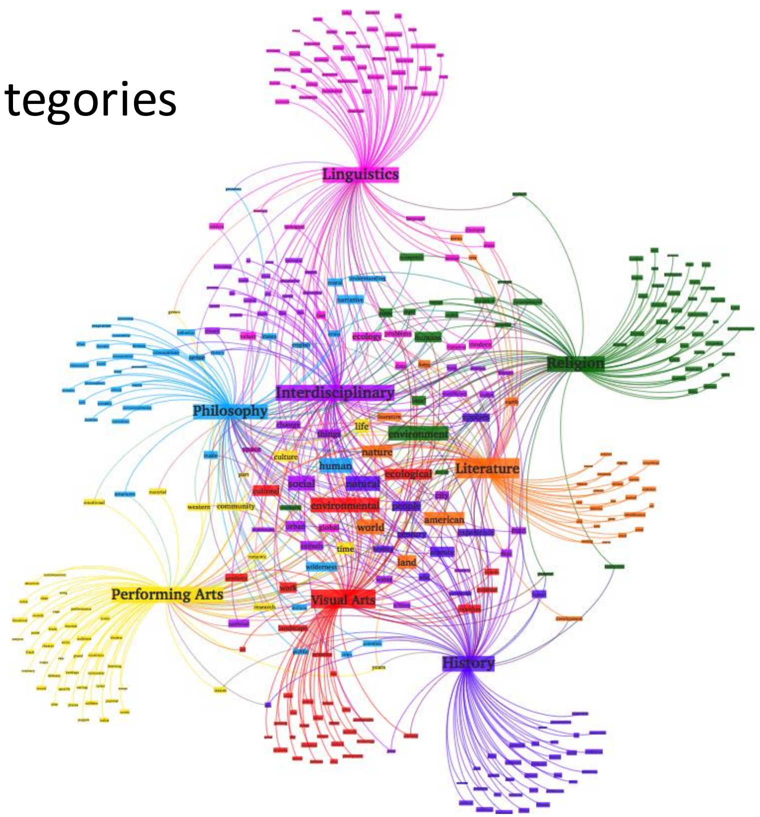
A Comprehensive Overview of Topic Modelling

Motivation

Motivation 1:

Given a corpus of text documents understand high level latent structure

- Organize the documents into thematic categories
- Find relationship between categories
- Representation learning of texts
- Global dependency identification
- Soft clustering of texts



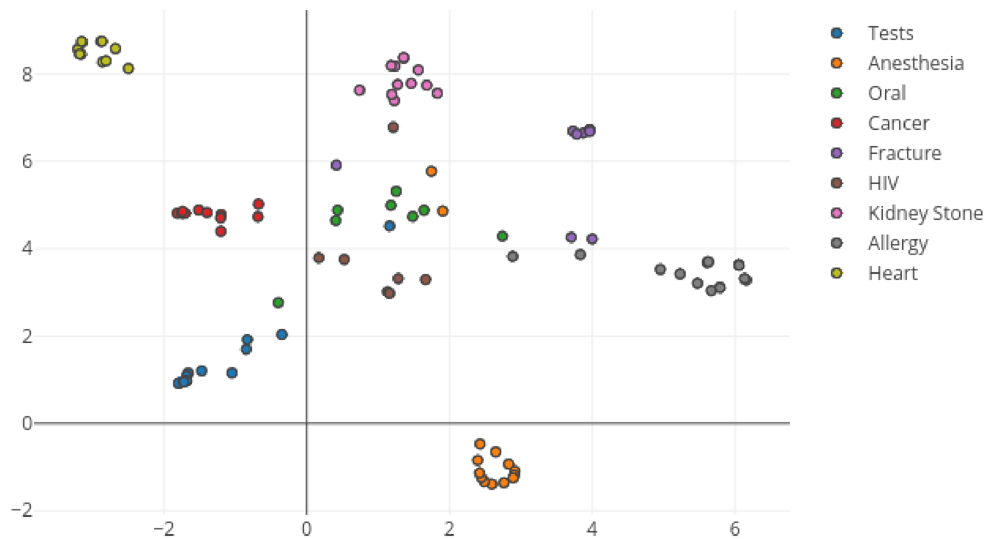
Ref - https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05

Motivation

Motivation 2:

Dimensionality reduction

- Convert sparse document-word matrix into low dimensional matrix
- Better interpretation at a lower dimension



t-SNE plot on low dimensional representation of patient discharge summaries

What is Topic Modelling

- An unsupervised text mining method
- Assumes -
  - Each text document is a mixture of latent (hidden) topics
  - Each topic is a collection of fixed set of words
- Fixed number of topics

Text = “The economy has crashed by 10%”

Text = 0.65 * Topic1 + 0.01 * Topic2 + 0.04 * Topic3 + 0.2 * Topic4 + 0.1 * Topic5

Topic1: 0.3 * finance + 0.25 * market + 0.15 * economy + 0.1 * stock + 0.06 * sector

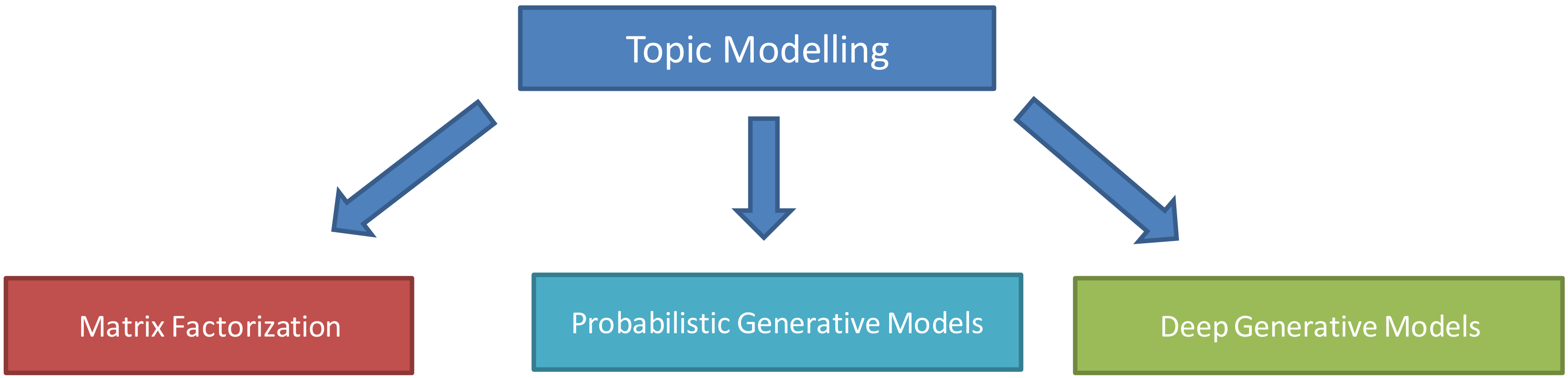
Topic2: 0.35 * crime + 0.22 * violence + 0.15 * war + 0.12 * gun + 0.09 * fire

Topic3: 0.4 * politic + 0.15 * globe + 0.1 * policy + 0.08 * government + 0.05 * ministry

Topic4: 0. 3 * accident + 0.2 * car + 0.16 * crash + 0.1 * driver + 0.05 * casualty

Topic5: 0.3 * bad + 0.3 * die + 0*15 * kill + 0.1 * virus + 0.1 * hospital

Topic Modelling Techniques



- NMF (non-negative MF)
- LSA (Latent Semantic Analysis)

- pLSA (Probabilistic Latent Semantic Analysis)
- LDA (Latent Dirichlet Allocation)

- GAN/VAE based
- Hybrid models

NMF based Topic Modelling

Input

- X : document-term matrix of size N x M
  - Count Matrix
  - Tf-idf Matrix

Parameters

- t : Number of topics (Latent dimension)

Outputs

- W : Document-topic matrix of size N x t
- H : Word-topic matrix of size M x t

such that

$$minimize \, \|X - WH\|_F^2 \; w.r.t. \, W, H \; s.t. \, W, H \geq 0$$

How to Solve NMF

- The loss function is minimized using iterative method (e.g. – SGD)

$$H \leftarrow H \odot \frac{W^T X}{W^T W H}$$

$$W \leftarrow W \odot \frac{X H^T}{W H H^T}$$

- Regularization can be added with the loss function

LSA (Landauer, T.K et al. 1998)

Input

- A : document-term matrix of size *N* x *M*
  - Count Matrix
  - Tf-idf Matrix

Parameters

- t : Number of topics (Latent dimension)

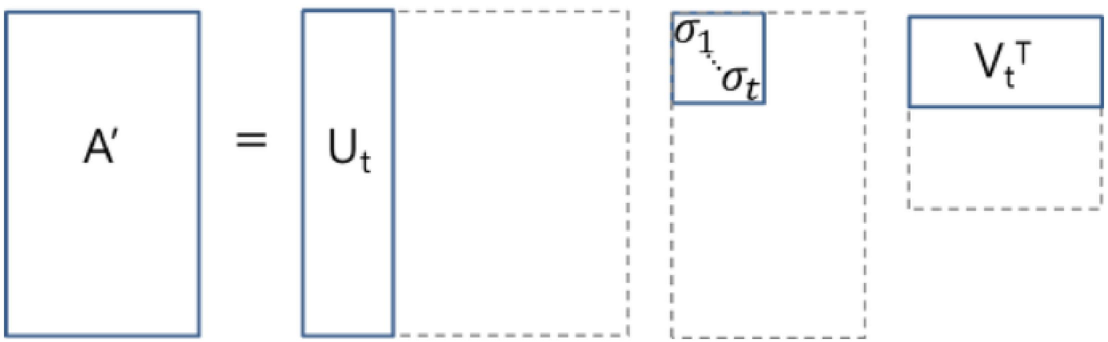
Outputs

$$A \approx U_t S_t V_t^T$$

- U_t: Document-topic matrix of size *N* x *t*
- S_t: Matrix with singular values of *A* of size *t* x *t*
- V_t: Word-topic matrix of size *M* x *t*

Solving LSA

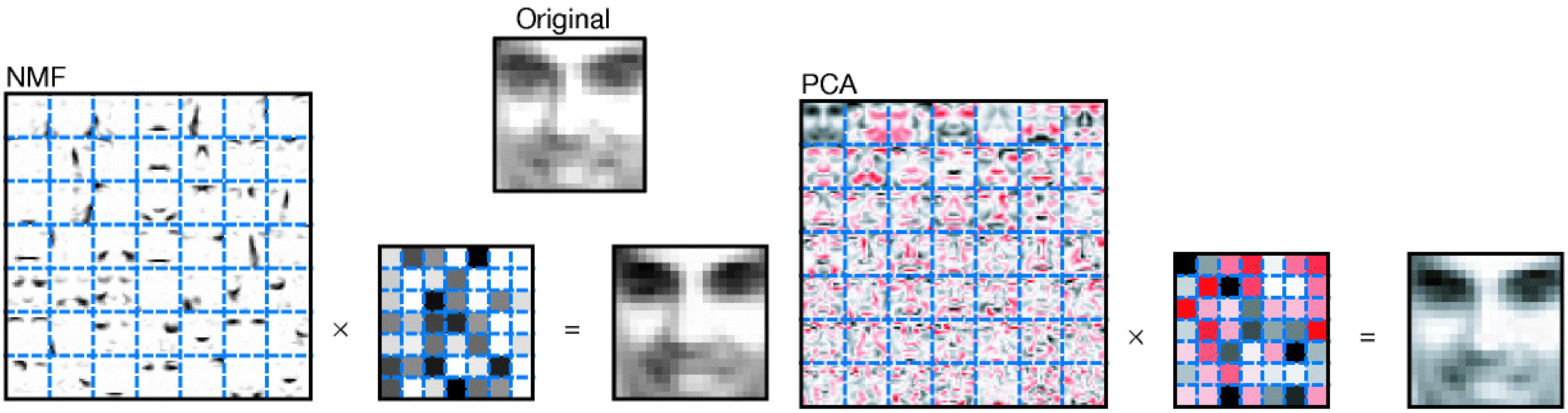
- Singular Value Decomposition of document-term matrix
- Pick *t* most significant dimensions



Matrix based Topic Modelling

Question: Why NMF is preferred over SVD/PCA?

Negative components are difficult to interpret



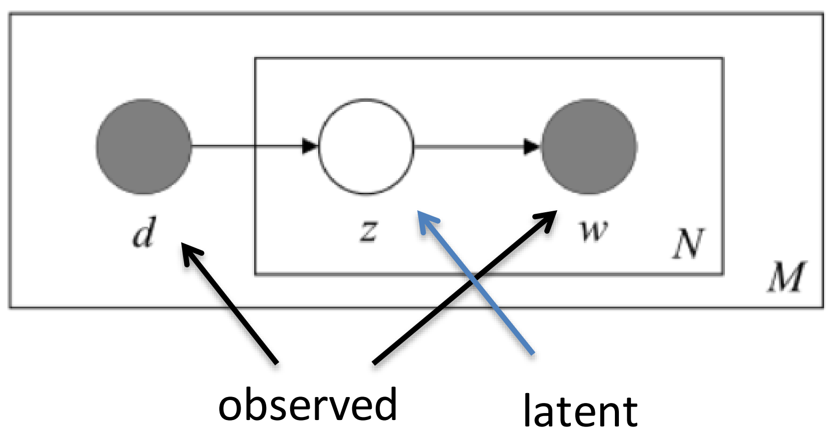
Ref - https://www.nature.com/articles/44565

pLSA (Hofmann 1999)

- Use probabilistic method instead of SVD
- Generative model for *P(d,w)* for each document *d* and word *w*

Model assumption

- Given a document d, topic z is present with $P(z|d)$
- Given a topic z, word w is drawn with $P(w|z)$



Solving pLSA

For each document *d* and word *w*,

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

$$U_t S_t V_t^T$$

Solve using Expectation-Maximization

E Step:

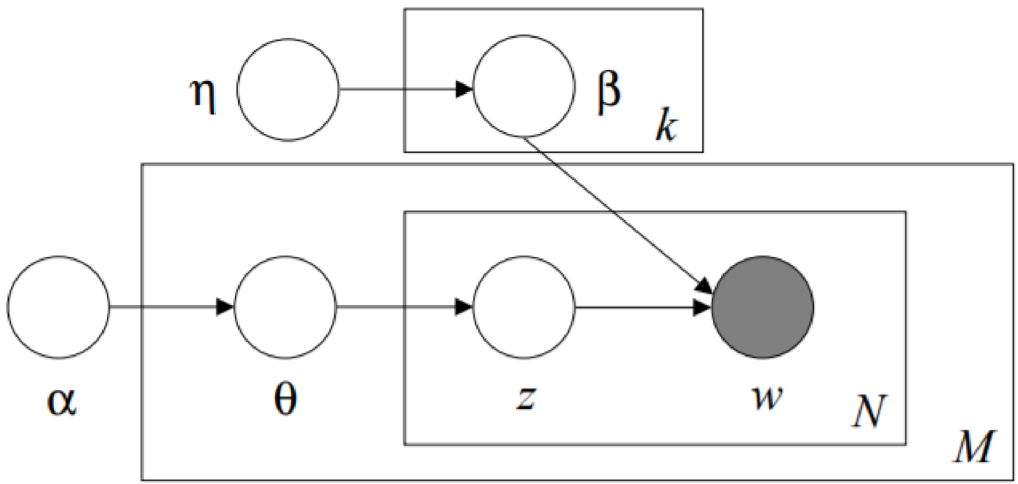
$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z' \in \mathcal{Z}} P(z')P(d|z')P(w|z')}$$

M step:

$$\begin{aligned}
P(w|z) &\propto \sum_{d \in \mathcal{D}} n(d,w)P(z|d,w), \\
P(d|z) &\propto \sum_{w \in \mathcal{W}} n(d,w)P(z|d,w), \\
P(z) &\propto \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} n(d,w)P(z|d,w).
\end{aligned}$$

LDA (Blei et al. 2003)

- pLSA is not well-defined generative model, as there is no natural way to assign probability to unseen document
- LDA is a Bayesian version of pLSA

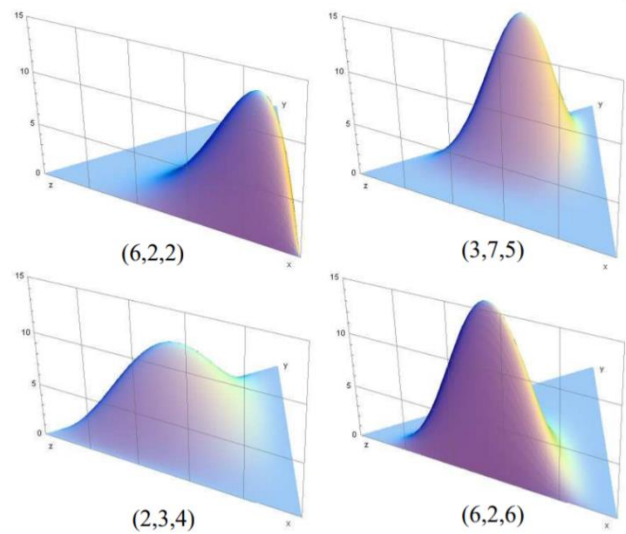


- α,η are Dirichlet hyper parameters
- θ,β are drawn from Dir(α) and Dir(η)
- z is drawn from Mult(θ)

Dirichlet Distribution

- Dirichlet distribution can be thought as a distribution over probability simplex.
- It is conjugate prior for Multinomial distribution (posterior distribution same as prior)

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{d} \alpha_i)}{\prod_{i=1}^{d} \Gamma(\alpha_i)} \prod_{i=1}^{d} x_i^{\alpha_i - 1}; \quad \text{for "observations"}: \sum_{i=1}^{d} x_i = 1, \quad x_i \geq 0$$



Generative Process of LDA

For each topic $k \in \{1, \ldots, K)$:

Sample $\beta_k \sim Dir(\eta)$

For each document $d \in \{1, \ldots M\}$:

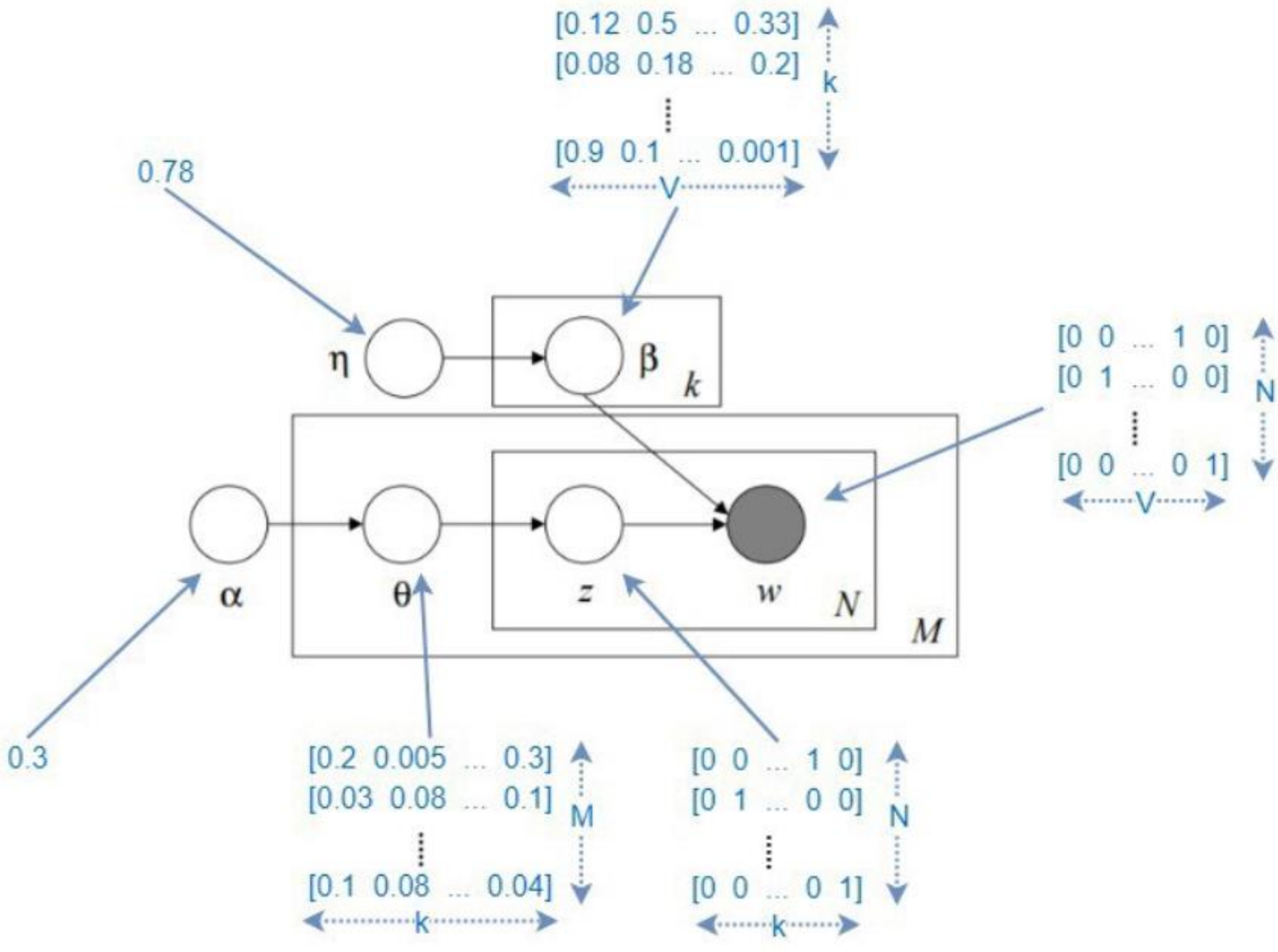
Sample $\theta_d \sim Dir(\alpha)$

For each word $n \in \{1, \ldots, N_d)$:

Sample $z_{dn} \sim Mult(1, \theta_d)$

Sample $x_{dn} \sim Mult(1, \beta_k)$

Generative Process of LDA



Ref - https://medium.com/@souravboss.bose/comprehensive-topic-modelling-with-nmf-lsa-plsa-lda-lda2vec-part-2-e3921e712f11

Posterior Calculation for LDA

For each word *w* in document *d,*

$$p(\theta, \beta, \boldsymbol{z}, \boldsymbol{w} \,|\alpha, \eta) = p(\theta|\alpha)p(\beta|\eta)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta)$$

Integrate out latent variable to calculate likelihood -

$$p(\boldsymbol{w} \,|\alpha, \eta) = \iint p(\theta|\alpha)p(\beta|\eta)\prod_{n=1}^{N}\sum_{z_n} p\,(z_n|\theta)p(w_n|z_n, \beta)d\theta d\beta$$

Finally, calculate posterior -

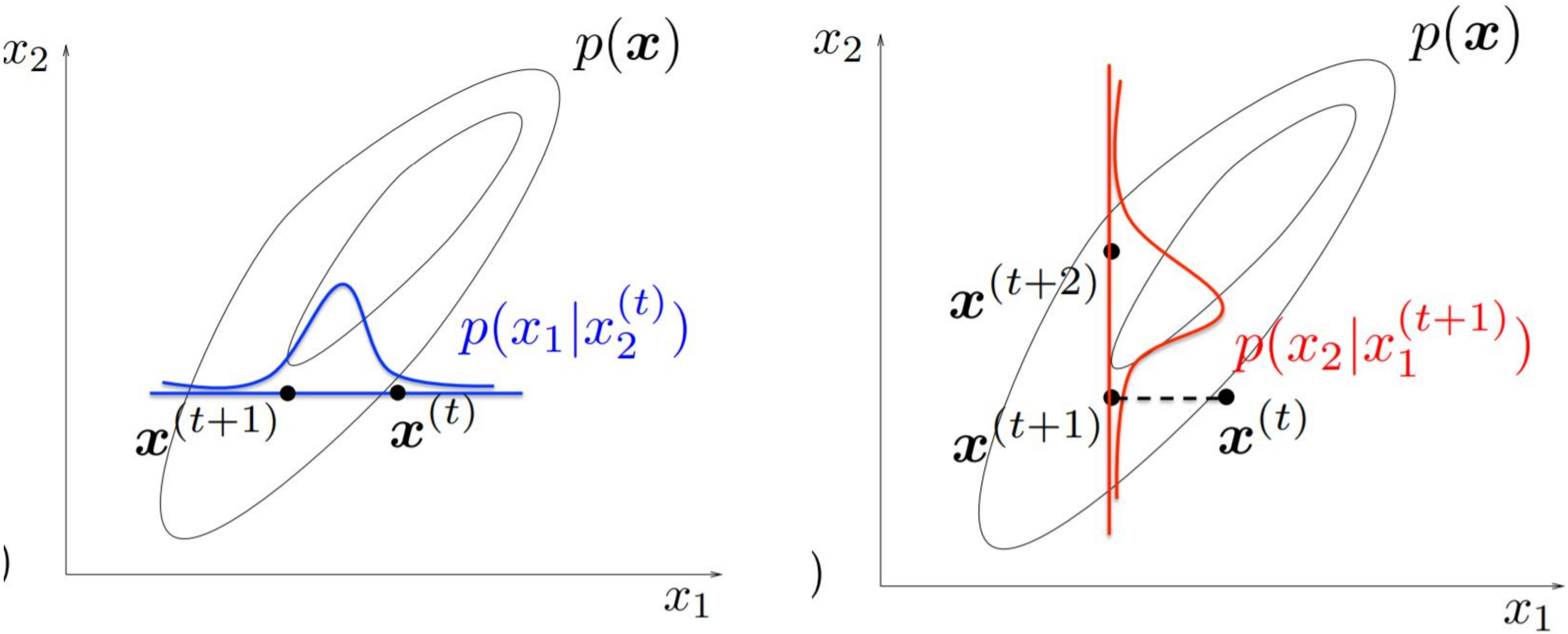
$$p(\theta, \beta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \eta) = \frac{p(\theta, \beta, \boldsymbol{z}, \boldsymbol{w} \,|\alpha, \eta)}{p(\boldsymbol{w} \,|\alpha, \eta)}$$

Posterior is intractable!! Need approximate inference to solve.

Inference for LDA

1. Gibbs Sampling – Gibbs sampling is a method of a Markov Chain Monte Carlo (MCMC). Iteratively sample for a variable by keeping all others fixed
2. Variational Inference – Solve an intractable posterior with a tractable distribution

Gibbs Sampling



Variational Inference

Approximate the intractable posterior $p(H|D)$ with a tractable distribution $q(H|D,V)$, where V is a set of free variational parameter

Variational Inference:

Find the free parameters V which minimizes KL divergence $KL(q||p)$

Gibbs Sampling for LDA

For each word *w* in document *d,* at ith iteration of Collapsed Gibbs sampling

$$p(z_{d,n} = k | \boldsymbol{z}_{-d,n}, \alpha, \eta) = \underbrace{\frac{N_{d,k} + \alpha}{\sum_i N_{d,i} + K\alpha}}_{\theta_{d,k}} \underbrace{\frac{V_{k,w} + \eta}{\sum_w V_{k,w} + |V|\eta}}_{\beta_{k,w}}$$

Variational Inference for LDA

$$\arg \min_{\vec{\gamma}_{1:D}, \vec{\lambda}_{1:K}, \vec{\phi}_{1:D,1:N}} \mathrm{KL}(q(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K}) || p(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K} \mid w_{1:D,1:N}))$$

Loss function to be optimized -

$$\mathcal{L} = \sum_{k=1}^{K} \mathrm{E}[\log p(\vec{\beta}_k \mid \eta)] + \sum_{d=1}^{D} \mathrm{E}[\log p(\vec{\theta}_d \mid \vec{\alpha})] + \sum_{d=1}^{D} \sum_{n=1}^{N} \mathrm{E}[\log p(Z_{d,n} \mid \vec{\theta}_d)] + \sum_{d=1}^{D} \sum_{n=1}^{N} \mathrm{E}[\log p(w_{d,n} \mid Z_{d,n}, \vec{\beta}_{1:K})] + \mathrm{H}(q),$$

One iteration of mean field variational inference for LDA

1) For each topic k and term v:

$$\lambda_{k,v}^{(t+1)} = \eta + \sum_{d=1}^{D} \sum_{n=1}^{N} 1(w_{d,n} = v)\phi_{n,k}^{(t)}.$$

2) For each document d:

   (a) Update γ_d:

$$\gamma_{d,k}^{(t+1)} = \alpha_k + \textstyle\sum_{n=1}^{N} \phi_{d,n,k}^{(t)}.$$

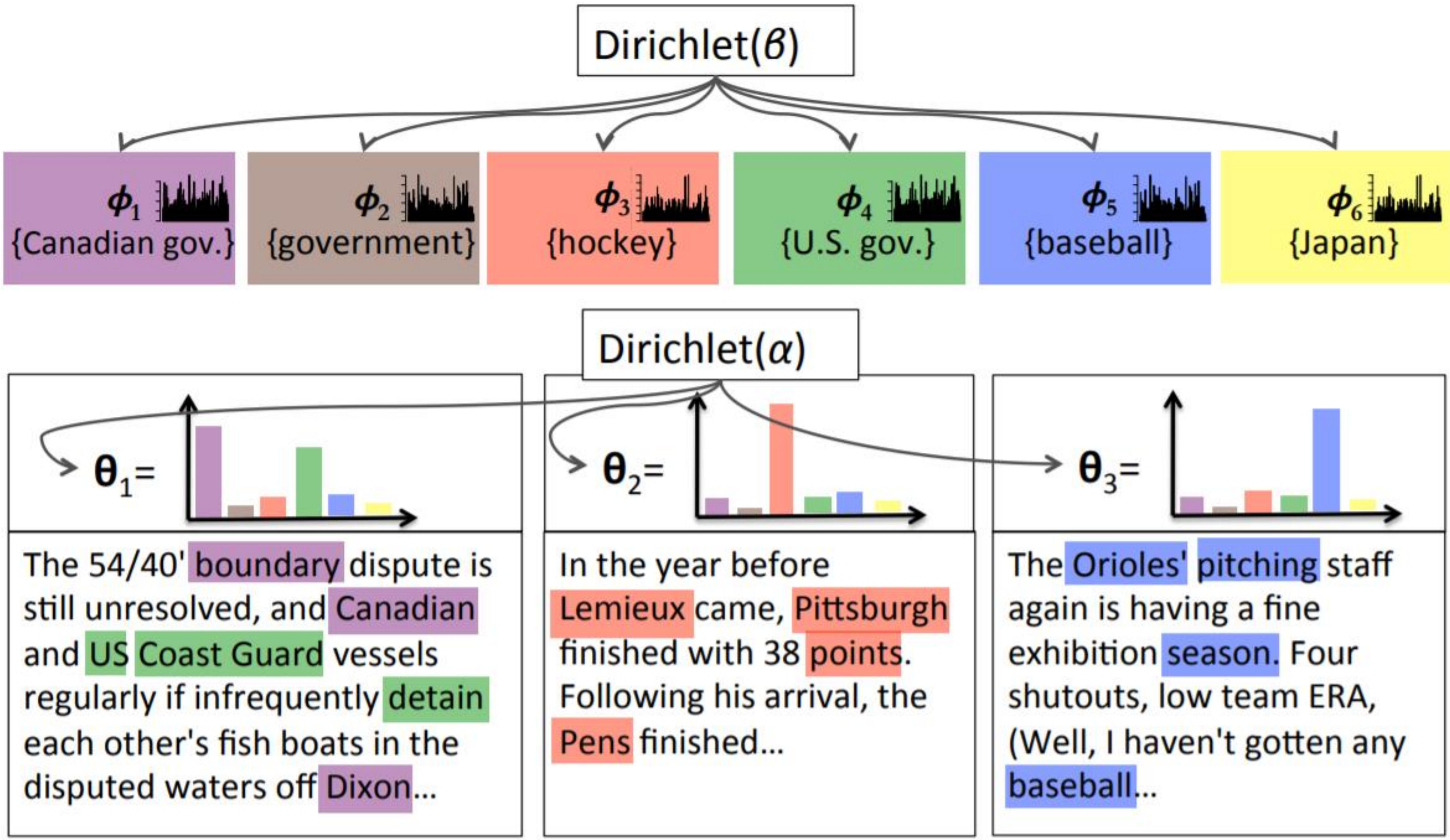
   (b) For each word n, update $\vec{\phi}_{d,n}$:

$$\phi_{d,n,k}^{(t+1)} \propto \exp\left\{\Psi(\gamma_{d,k}^{(t+1)}) + \Psi(\lambda_{k,w_n}^{(t+1)}) - \Psi(\textstyle\sum_{v=1}^{V} \lambda_{k,v}^{(t+1)})\right\},$$

   where Ψ is the digamma function, the first derivative of the $\log \Gamma$ function.

Ref - https://courses.engr.illinois.edu/cs598jhm/sp2010/Slides/Lecture07HO.pdf

LDA Illustration



Ref - https://www.cs.cmu.edu/~mgormley/courses/10701-f16/slides/lecture20-topic-models.pdf

Recent Works based on LDA

- Lda2vec (Chris Moody 2016) – Add word vector with topic document vector to predict context word
- Biterm Topic Model for Short Texts (Yan et. al. 2013) – Instead of unigram, used biterms to tackle short text sparsity
- Topic Modelling with Word Embeddings (Qiang et.al. 2016)
- Author-Topic model for Authors and Documents (Rosen-Zvi et. al. 2012)

Deep learning based Topic Models

- Autoencoding Variational Inference for Topic Models (Srivastava, Sutton 2017)
- ATM: Adversarial-neural Topic Model (Wang et. al. 2019)
- Topic Modelling with Wasserstein Autoencoders (Nan et. al. 2019)

Evaluation of Topic Models

- **Perplexity** – Normalized log-likelihood of held out test data

$$per(D_{test}) = exp\{-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\}$$

However, perplexity may not yield human interpretable topics

- **Topic coherence –** Coherence is a score to measure degree of semantic similarity between high scoring words in topic.

$$CoherenceScore = \sum_{i<j} score(w_i, w_j)$$

1. Extrinsic UCI measure:

$$SCORE_{UCI}(w_i, w_j) = log\frac{p(w_i, w_j)}{p(w_i)P(w_j)}$$

2. Intrinsic UMass measure:

$$SCORE_{UMass}(w_i, w_j) = log\frac{D(w_i, w_j)+1}{D(w_i)}$$

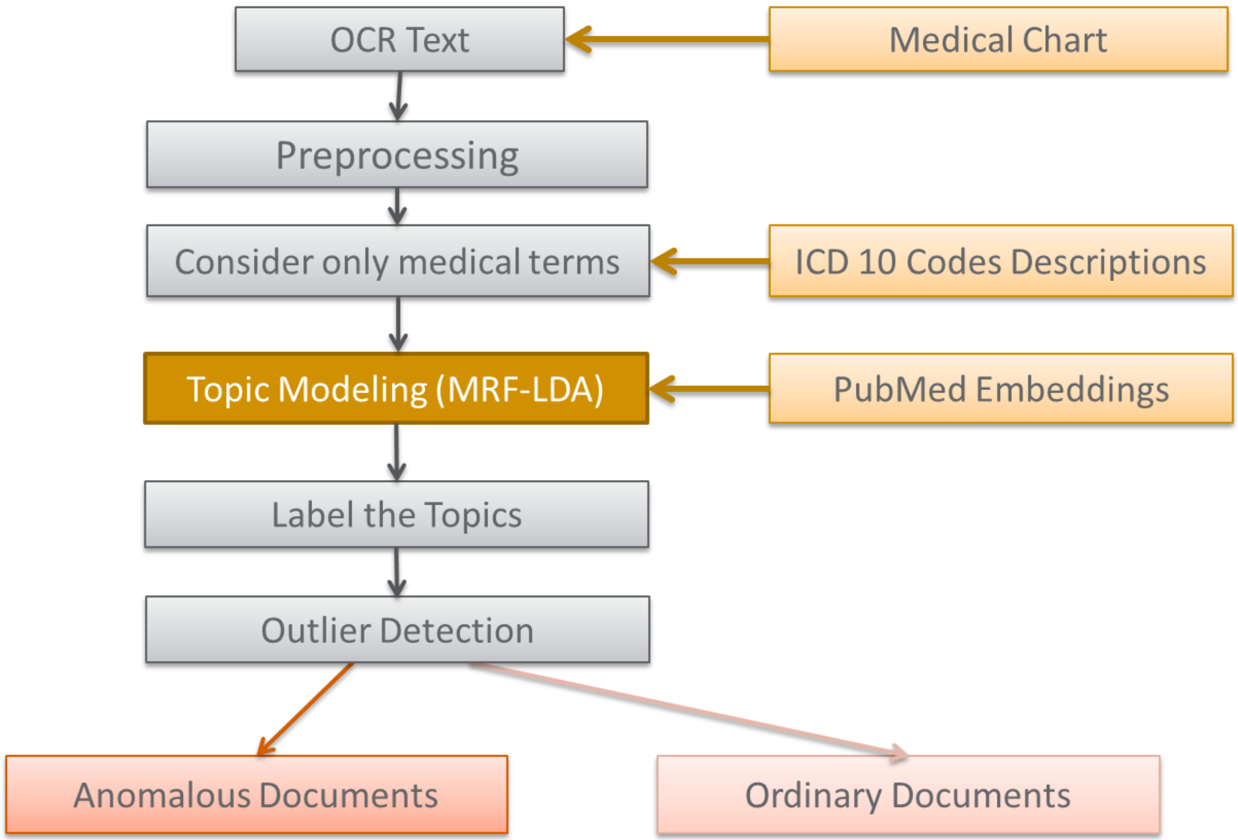
Evaluation of Topic Models

- **Clustering based methods –**
  - Silhouette Score
  - NMI (normalized mutual info)

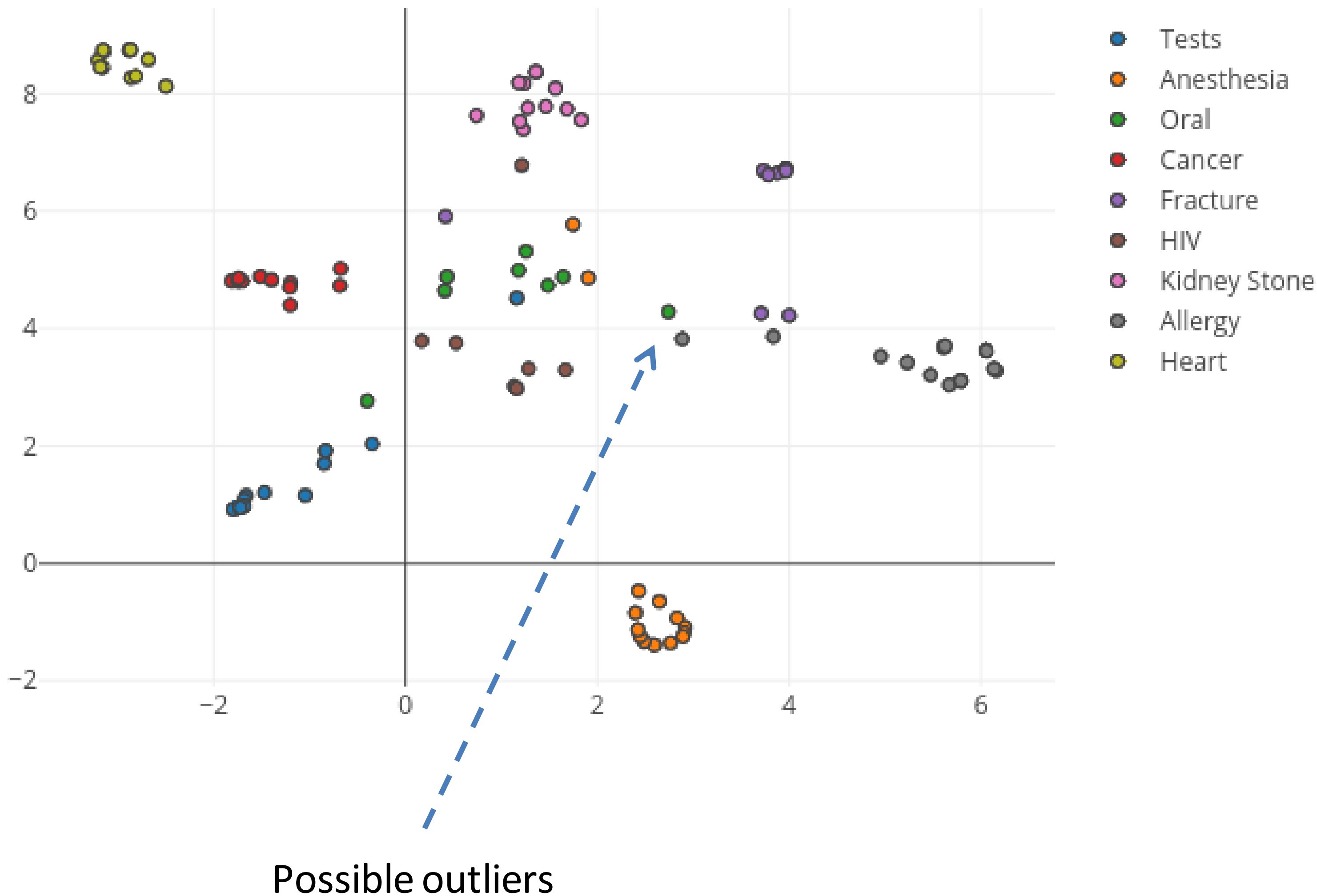
- **Human evaluation (Qualitative methods)**

Other Applications of Topic Modelling

- Outlier detection



Outlier Detection with Topic Modelling



Other Applications of Topic Modelling

- Spacial LDA (Wang & Grimson, 2007)