Topic Modelling (L101)

Ann Copestake

Computer Laboratory
University of Cambridge

October 2016

Outline of today's lecture (and some of next week's?)

- Clustering
- Brief overview of LDA
- Latent variables
- Dirichlet distribution
- Gibbs sampling (lecture 6)
- Hyperparameters
- LDA again

NB: some slides borrowed from Diarmuid Ó Séaghdha
`http://www.cl.cam.ac.uk/~do242/Teaching/HIT-MSRA-2011/`

Outline.

Clustering vs classification

- Classification: predefined classes, usually supervised.
- Clustering: groupings induced from data:
  - Hard clustering: each item is in one cluster
  - Soft clustering: probability distribution over clusters
- Number of clusters known in advance? Required for e.g., k-means, but potentially useful if open.
- Probabilistic soft clustering: formally very like generative classifier, but no labels observed when training!
- Clustering is good in contexts where categories are (somewhat) arbitrary.
- Downsides: labelling clusters may be hard/impossible: difficult to evaluate as standalone system.

Word Sense Disambiguation vs Word Sense Induction

- WSD: different tokens classified according to senses defined in a particular dictionary (e.g., Wordnet).
  BUT:
  - Dictionaries differ!
  - Lexicographers must distinguish between senses, but don't actually see them as hard classes (apart from homonyms).
  - Utility of distinctions depends on the NLP application.
- WSI: different tokens clustered by similarity of use.
  - Something analogous to task-specific WSI happens implicitly in various NLP systems, but is not treated as a separate module.

Topic modelling (D)

Microsoft revenues hit a record as Xbox sales soar

The US technology giant Microsoft said its annual revenues hit a record of $69.94bn (£43.4bn).

Sales of the company's Xbox 360 videogame console and its Office software helped fuel the growth.

Net income at the world's biggest software maker jumped 23% to 23.15bn for the year.

The figures, which beat forecasts, showed final quarter revenues reached a record high of $17.37bn, leading to profits of $5.87bn.

Microsoft's business division, which includes Office software - is its biggest seller

Topic 1 “business”	Topic 2 “technology”	Topic 3 “travel”
0.7	0.3	0

Easyjet raises profits forecast

Shares in budget airline Easyjet have risen 18% after it raised its profit guidance for the year.

Revenue in the three months to June was up 23% on a year ago to £935m after it increased its number of flights.

The firm said its new strategy, which includes appealing to more business customers, was seeing "good progress", with a 20% increase in business passengers seen in the quarter.

The airline now expects a full-year profit of between £200m and £230m.

Analysts had forecast a £179m profit.

Easyjet said passenger numbers had increased 17.3% compared with the same quarter a year earlier.

Topic 1 “business”	Topic 2 “technology”	Topic 3 “travel”
0.6	0	0.4

Topic classification vs topic modelling

- Predefined document classes are sometimes available (libraries, scientific abstracts).
- But classes are often not predefined, predefined classes may be unsuitable.
- Automatically induced topics give an alternative way of organising big data collections.
- But: topic modelling (usually) doesn't give a very good approximation to conventional categories.
- However: this doesn't matter for many applications ...

Example - biomedical corpus (D)

- Most probable words for topics found by LDA in the OpenPMC corpus of biomedical scientific articles:

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
sequence	exposure	important	retinal	neurons
sequences	levels	number	lens	mice
genome	study	large	cells	receptor
genes	health	specific	retina	receptors
gene	data	studies	patients	pain
protein	environmental	result	expression	rats
species	risk	potential	corneal	synaptic
dna	effects	type	eye	brain
data	children	represent	rpe	nerve
proteins	studies	long	mutation	neuronal

Example - Twitter corpus (D)

- Most probable words for topics found by LDA in a corpus of Twitter users in London during 2010:

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
world	lol	blog	love	baby
cup	haha	post	#xfactor	kids
england	good	updated	factor	family
#worldcup	dont	comment	big	children
football	yeah	published	cheryl	school
south	hey	entry	amazing	child
spain	love	blogs	show	parents
africa	hope	blogging	live	fun
game	gonna	posts	john	great
germany	time	posting	brother	toys

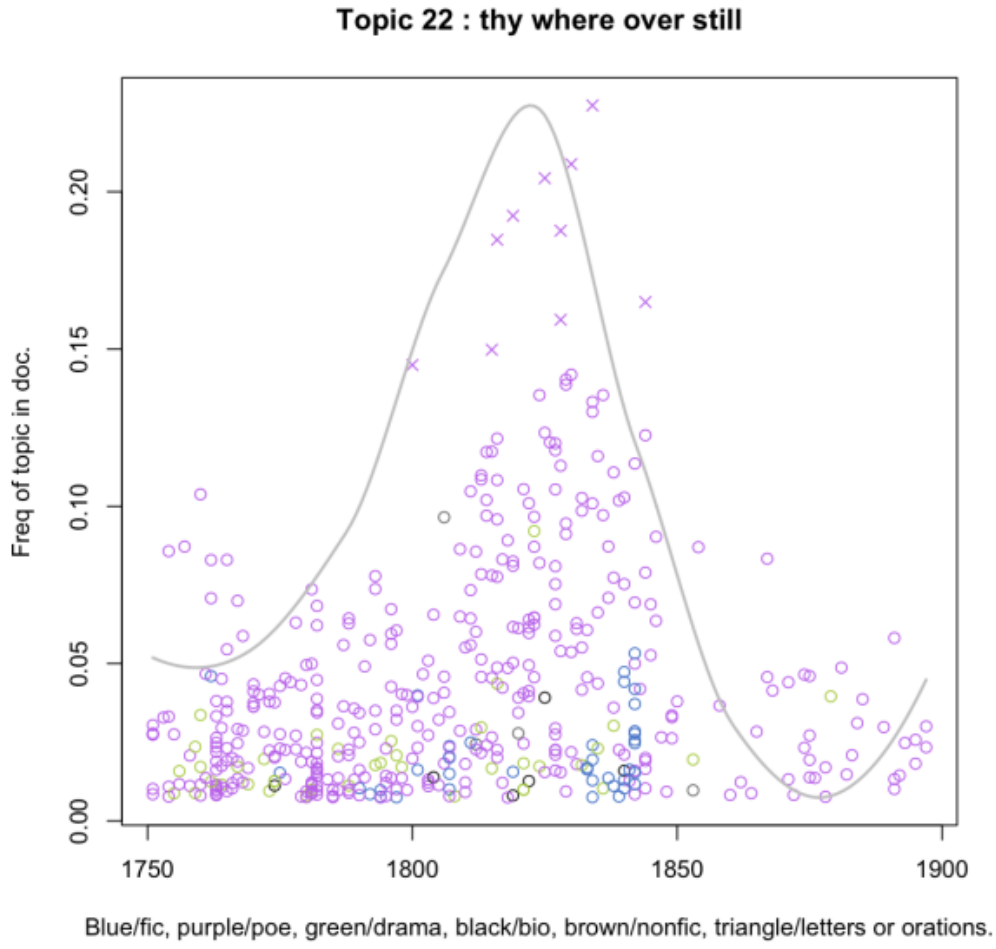
But what are topics?

- LDA is almost invariably illustrated with clean subject topics (as earlier), but low frequency topics are usually much less coherent.
- Topic modelling has become very popular within digital humanities: e.g., `http://dsl.richmond.edu/dispatch/pages/intro`
- Historians are interested in topics like TRADE, but other scholars find less obvious topics useful.
- From now on, implicit scare quotes around the term “topic”!

```
TOPIC 22 : thy where over still when oh deep bright wild eye yet light tis whose brow each round
through many dark wave beneath twas around hour like while away thine those page hath lone sky
spirit song oft notes home mid grave vain again though far mountain shore soul ocean night
OF 150 TOPICS this is # 9 in desc order, with 491728 words. Related topics:
```

"A topic like this one is hard to interpret. But for a literary scholar, that's a plus."

https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/



https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Outline.

Documents as mixtures of topics (D)

Microsoft revenues hit a record as Xbox sales soar

The US technology giant Microsoft said its annual revenues hit a record of $69.94bn (£43.4bn).

Sales of the company's Xbox 360 videogame console and its Office software helped fuel the growth.

Net income at the world's biggest software maker jumped 23% to 23.15bn for the year.

The figures, which beat forecasts, showed final quarter revenues reached a record high of $17.37bn, leading to profits of $5.87bn.

Microsoft's business division, which includes Office software - is its biggest seller

Topic 1 “business”	Topic 2 “technology”	Topic 3 “travel”
0.7	0.3	0

Easyjet raises profits forecast

Shares in budget airline Easyjet have risen 18% after it raised its profit guidance for the year.

Revenue in the three months to June was up 23% on a year ago to £935m after it increased its number of flights.

The firm said its new strategy, which includes appealing to more business customers, was seeing "good progress", with a 20% increase in business passengers seen in the quarter.

The airline now expects a full-year profit of between £200m and £230m.

Analysts had forecast a £179m profit.

Easyjet said passenger numbers had increased 17.3% compared with the same quarter a year earlier.

Topic 1 “business”	Topic 2 “technology”	Topic 3 “travel”
0.6	0	0.4

Documents as mixtures of topics (D)

Microsoft **revenues hit** a **record** as **Xbox** **sales soar**
Microsoft's business division which includes Office software is its biggest seller. Sales of the company's Xbox 360 videogame console and its Office software helped fuel the growth. Net income at the world's biggest software maker jumped 23% to 23.15bn for the year.

"business" "technology" "general"

Latent Dirichlet Allocation (LDA)

- LDA: latent variable model: a set of latent variables Z connects documents and words.
- In topic modelling terms: LDA clusters words (types) into topics and assigns documents a distribution over topics.
- Mixture model assumption: each word (token) in a document is associated with a single mixture component (i.e., topic).
- LDA can also be seen as a distributional model with dimensionality reduction (to $|Z|$-dimensional space)
- LDA can be used for distributional similarity and modelling selectional preferences.
- LDA related to LSA, but probabilistic rather than geometric notion of dimensionality reduction.

LDA as a generative model (intuitively)

LDA models documents as though they were generated by the following process:

1. Generate the topics (a topic is just a distribution over a fixed vocabulary)
2. Generate each document as follows:
   - 2.1 Randomly choose a distribution over topics
   - 2.2 To choose each word in the document:
     - 2.2.1 Randomly choose a topic from the distribution created in Step 2.1
     - 2.2.2 Randomly choose a word from that topic's distribution over the vocabulary

But we don't know the topics, or the distributions associated with the topics, so we have to work this out from the documents.

LDA - the generative story (D)

for topic $z \in \{1 \ldots |Z|\}$ **do**
  $\Phi_z \sim Dirichlet(\beta)$
end for
for document $d \in \{1 \ldots |D|\}$ **do**
  $\theta_d \sim Dirichlet(\alpha)$
  for word $i \in d$ **do**
    $z_i \sim Multinomial(\theta_d)$
    $w_i \sim Multinomial(\Phi_{z_i})$
  end for
end for

LDA - the maths (from Diarmuid's slides) I

- The joint distribution of observed and hidden variables is:

$$P(D, \mathbf{z}, \Phi, \theta; \alpha, \beta) = \prod_{d \in D} p(\theta_d; \alpha) \prod_{i \in d} P(z_i; \theta_d) P(w_i; \Phi_z)$$

- We can integrate out the multinomial parameters Φ and θ due to Dirichlet-multinomial conjugacy. This means we average over all possible parameter values rather than committing to one particular value.
- The posterior distribution over topic assignments $\mathbf{z}$ is:

$$\begin{aligned} P(\mathbf{z}|D; \alpha, \beta) \propto & P(D|\mathbf{z}; \beta) P(\mathbf{z}; \alpha) \\ & = \int_{\Phi} p(\Phi; \beta) \int_{\theta} p(\theta; \alpha) \cdot \\ & \qquad \prod_{d \in D} \prod_{i \in d} P(w_i|z_i; \Phi) P(z_i|d; \theta) \mathrm{d}\theta \mathrm{d}\Phi \end{aligned}$$

LDA - the maths (from Diarmuid's slides) II

- `http://www.isi.edu/natural-language/people/bayes-with-tears.pdf`
- Optimising the posterior distribution is an intractable problem; we must use approximate methods: for instance, Gibbs sampling.

Unpicking LDA

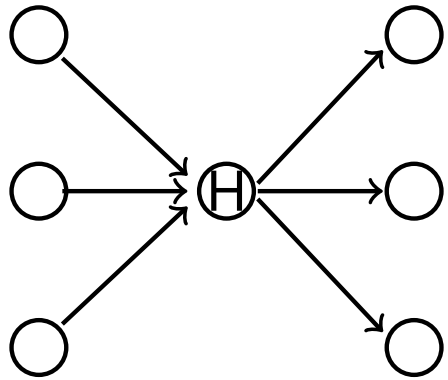
Various concepts used in LDA may well be unfamiliar:

- Latent variables
- Dirichlet distribution
- Gibbs sampling
- Hyperparameters ($|Z|$, α, β)

These are important in various ML algorithms. In the remainder of the lecture, I discuss these each (briefly) before returning to LDA.

Outline.

Latent variables in probabilistic approaches



- Latent (hidden) variables inserted between conditionally dependent observed variables: models have far fewer parameters.
- Latent variables can compute a compressed representation of the data.
- Induce some hidden structure, which we can examine.

Probabilistic models with latent variables

- Usually we have many more observed variables (e.g., documents, words) than latent variables (e.g., topics).
- The number of latent variables is a hyperparameter.
- Include many popular approaches, both continuous and discrete: Gaussian mixture model (GMM) etc.
- Often compute maximum likelihood (ML) or the maximum a posteriori (MAP) by expectation maximization (EM).
- k-means is a variant of EM for GMMs.
- Chapter 9 of Bishop (2006)

Outline.

Dirichlet distribution

Dirichlet is to the multinomial as the beta distribution is to the binomial.
Beta distribution:

- Suppose we have a coin which may or may not be fair: how do we estimate the probability of it coming up heads (i.e., estimate what the binomial actually is)?
- Suppose we toss it 4 times and it comes up heads every time: do we assume P(heads)$=1$?
- Probably not, because we know something about coins: we want to use this prior knowledge about the distribution.
- The beta distribution describes possible binomial distributions: the prior allows smoothing.
- Equivalent to pretending we've already done some coin tosses.

Priors (Bishop, 2006: p71)

- conjugate priors: priors are expressed to be in the same functional form as the posterior distribution, hence we can integrate out the parameters in LDA.
- Posterior distribution can become the prior for new experiments.
- add-one smoothing is just a special case (which often works quite well, so may be sensible if you really don't know anything much about the distribution to start with).

Dirichlet and Dirichlet priors

- Dirichlet is for the multinomial, so Dirichlet priors may be uniform (single value) or non-uniform (vector).
- In LDA experiments, the prior for the topic distribution (α) is usually treated as non-uniform (which is possible because the number of topics is usually fairly small) while β is treated as uniform.
- Experiments and discussion in Wallach et al (2009), but we'll treat them both as uniform here.
  Diarmuid's Gibbs sampling slides (below) have $\alpha = (1, 1)$: pretend it's just 1.

Outline.

Sampling in general

- Exact inference is intractable for most probabilistic methods.
- Monte Carlo methods are a way of making approximations via random samples.
- Markov Chain Monte Carlo methods: assume that the sample of one state is based on the prior state.
- Sampling itself has to be tractable (approximate the sampling procedure . . .)
- Bishop (2006: Chapter 11) discusses sampling in detail, including Gibbs Sampling.

LDA - the maths: a recap

The posterior distribution over topic assignments $\mathbf{z}$ is:

$$
\begin{aligned}
P(\mathbf{z}|D;\boldsymbol{\alpha},\boldsymbol{\beta}) \propto & P(D|\mathbf{z};\boldsymbol{\beta})P(\mathbf{z};\boldsymbol{\alpha}) \\
= & \int_{\boldsymbol{\Phi}} p(\boldsymbol{\Phi};\boldsymbol{\beta}) \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta};\boldsymbol{\alpha}) \cdot \\
& \prod_{d \in D} \prod_{i \in d} P(w_i|z_i;\boldsymbol{\Phi}) P(z_i|d;\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\Phi}
\end{aligned}
$$

Optimising the posterior distribution is an intractable problem; we must use approximate methods: for instance, Gibbs sampling.

Gibbs sampling for LDA (all from Diarmuid's slides) I

- Gibbs sampling is a general Markov Chain Monte Carlo method for evaluating probability distributions that are difficult or impossible to evaluate analytically; see `https://www.umiacs.umd.edu/~resnik/pubs/LAMP-TR-153.pdf` for an NLP-friendly introduction.
- The intuitive idea behind Gibbs sampling is to iterate through the dataset, updating one "small part" of the model at a time.
- Each update is non-deterministic: even from the same starting state two sampling runs will visit different states.
- Typically we run the sampler for a large number of iterations (e.g., 1000 passes through the corpus) and estimate the posterior using the final sampling state or an average over non-adjacent sampling states.

Gibbs sampling for LDA (all from Diarmuid's slides) II

- For LDA, each sampling iteration updates the topic assignment z_i of each token w_i in the corpus in succession, fixing the assignments of all other tokens and using those assignments to compute the distribution over values of z_i:

$$P(z_i = z | w_i, \mathbf{z}^{-i}, \mathbf{w}^{-i}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{f_{zd_i}^{-i} + \alpha_z}{f_d^{-i} + \sum_{z'} \alpha_{z'}} \frac{f_{zw_i}^{-i} + \beta_{w_i}}{f_z^{-i} + \sum_{w'} \beta_{w'}}$$

where we use the following notation:

$\mathbf{w}^{-i}$	all words other than the ith token
$\mathbf{z}^{-i}$	all topic assignments other than the ith token
f_{zd}	number of tokens in document d assigned to topic z
f_{zw}	number of tokens of type w in the corpus assigned to z
f_d	length in tokens of document d
f_z	number of tokens assigned to topic z

Gibbs sampling pseudocode – outer loop

Given documents D, topic vocabulary Z, no. of iterations *ITS*, hyperparameters α, β:

for $d = 1$ **to** $|D|$ **do**
  for all $w \in d$ **do**
    TopicAssignments[z] = RandomInt($|Z|$)
  end for
end for
for $i = 1$ **to** *ITERATIONS* **do**
  for $d = 1$ **to** $|D|$ **do**
    for all $w \in d$ **do**
      InnerLoop(w,d,*TopicAssignments*)
    end for
  end for
end for

Gibbs sampling pseudocode – outer loop

Given documents D, topic vocabulary Z, no. of iterations ITS, hyperparameters α, β:

for $d = 1$ **to** $|D|$ **do**
  for all $w \in d$ **do**
    $TopicAssignments[z]$ = RandomInt($|Z|$)
  end for
end for
for $i = 1$ **to** $ITERATIONS$ **do**
  for $d = 1$ **to** $|D|$ **do**
    for all $w \in d$ **do**
      InnerLoop(w,d,$TopicAssignments$)
    end for
  end for
end for

Gibbs sampling pseudocode – inner loop

To update the topic assignment for a single w:

DecrementCounts(w) {Subtract 1 from all counts related to w}
for $z = 1$ **to** $|Z|$ **do**
  $Score[z]$ = ScoreTopic(w,d,z)
  Sum = Sum + $Score[z]$
end for
r = Random(Sum) {Random number between 0 and Sum}
$newZ$ = -1
while $r >= 0$ **do**
  $newZ$ = $newZ$ + 1
  r = r - $Score[z]$
end while
$TopicAssignments[w]$ = $newZ$
IncrementCounts(w,d,z)

Gibbs sampling example

Toy example: $|Z| = 2, \alpha = (1, 1), \beta = 0.1, |V| = 4$

Doc 1: a a b a b a a

Doc 2: c c d a c

Doc 3: a b b b

Doc 4: d

Gibbs sampling example

Toy example: $|Z| = 2, \alpha = (1, 1), \beta = 0.1, |V| = 4$

Doc 1:	a	a	b	a	b	a	a
	2	1	2	1	2	1	1
Doc 2:	c	c	d	a	c		
	1	2	2	1	2		
Doc 3:	a	b	b	b			
	2	2	1	2			
Doc 4:	d						
	1						

Random initialisation

Gibbs sampling example

Toy example: $|Z| = 2, \alpha = (1, 1), \beta = 0.1, |V| = 4$

Doc 1:	a	a	b	a	b	a	a
	???	1	2	1	2	1	1
Doc 2:	c	c	d	a	c		
	1	2	2	1	2		
Doc 3:	a	b	b	b			
	2	2	1	2			
Doc 4:	d						
	1						

$$P(z_1 = 1) \propto \frac{f_{z_1 d_1} + \alpha_1}{f_{d_1} + \sum_{z'} \alpha_{z'}} \frac{f_{z_1 w_a} + \beta_{w_a}}{f_{z_1} + \sum_{w'} \beta_{w'}}$$

$$P(z_1 = 2) \propto \frac{f_{z_2 d_1} + \alpha_2}{f_{d_1} + \sum_{z'} \alpha_{z'}} \frac{f_{z_2 w_a} + \beta_{w_a}}{f_{z_2} + \sum_{w'} \beta_{w'}}$$

Gibbs sampling example

Toy example: $|Z| = 2, \alpha = (1, 1), \beta = 0.1, |V| = 4$

Doc 1:	a	a	b	a	b	a	a
	???	1	2	1	2	1	1
Doc 2:	c	c	d	a	c		
	1	2	2	1	2		
Doc 3:	a	b	b	b			
	2	2	1	2			
Doc 4:	d						
	1						

$$P(z_1 = 1) \propto \left(\frac{4+1}{6+2}\right)\left(\frac{5+0.1}{8+0.4}\right)$$

$$P(z_1 = 2) \propto \left(\frac{2+1}{6+2}\right)\left(\frac{1+0.1}{8+0.4}\right)$$

Gibbs sampling example

Toy example: $|Z| = 2, \alpha = (1, 1), \beta = 0.1, |V| = 4$

Doc 1:	a	a	b	a	b	a	a
	???	1	2	1	2	1	1
Doc 2:	c	c	d	a	c		
	1	2	2	1	2		
Doc 3:	a	b	b	b			
	2	2	1	2			
Doc 4:	d						
	1						

$P(z_1 = 1) \propto 0.38 \qquad P(z_1 = 2) \propto 0.05$

Sample randomly in $(0, 0.43)$: 0.12

So z_1 is set to 1

Gibbs sampling example

Toy example: $|Z| = 2, \alpha = (1, 1), \beta = 0.1, |V| = 4$

Doc 1:	a	a	b	a	b	a	a
	1	???	2	1	2	1	1
Doc 2:	c	c	d	a	c		
	1	2	2	1	2		
Doc 3:	a	b	b	b			
	2	2	1	2			
Doc 4:	d						
	1						

And move on to the next token. . .

LDA - Interpreting the results

- The states visited by Gibbs sampler (after the burnin period) can be used to estimate various properties of interest.
- From a computational semantics perspective we are most interested in
  (a) Estimating the topic-word and document-topic distributions Φ and θ.
  (b) Evaluating the learned model in an application or through comparison to human judgements.

LDA - Interpreting the results

- The posterior mean of the topic distribution θ_d for a document d is given by:

$$\hat{\theta}_{dz(MEAN)} = \frac{f_{dz} + \alpha_z}{f_d + \sum'_z \alpha'_z}$$

- The posterior mean of the word distribution Φ_z for a topic z is given by:

$$\hat{\Phi}_{zw(MEAN)} = \frac{f_{zw} + \beta_w}{f_z + \sum'_w \beta'_w}$$

- Recall that the effect of the Dirichlet prior is to smooth the estimation of a multinomial distribution.

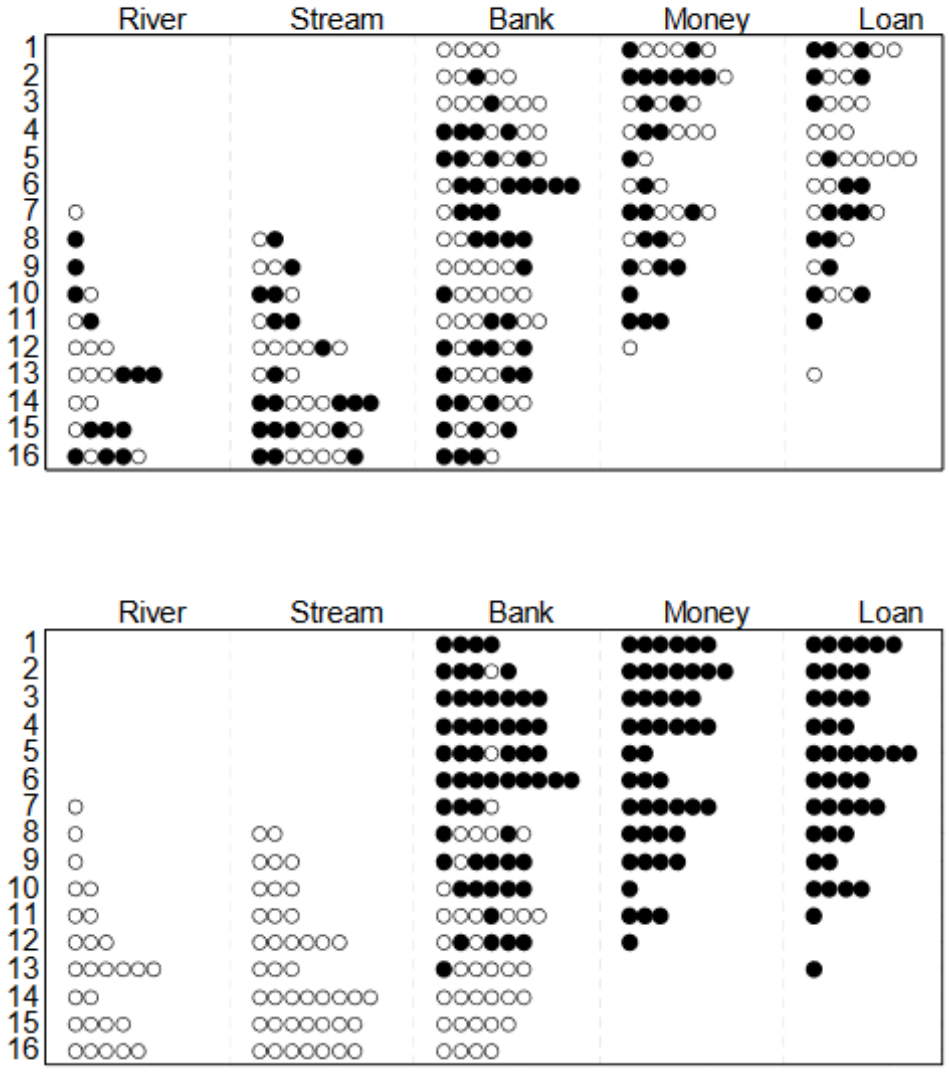


Figure 7. An example of the Gibbs sampling procedure.

From Stevyers and Griffiths (2007)

Outline.

Dirichlet priors in LDA

- α and β are 'officially' the way we give the model information about the distributions
- they smooth the estimation of the multinomial
- more practically: they allow us to experiment ('tune hyperparameters') until we get the desired results . . .

Hyperparameters

- General experimental methodology:
  - choose an evaluation metric (preferably not exactly the same as final evaluation)
  - explore parameter values on the development set to maximize the results
  - finally: evaluate on test set (be careful about repeating evaluations too often!)
- the parameter space to be explored can be very large ...
- number of latent variables $|Z|$ is also a hyperparameter for LDA, but mostly doesn't matter as long as it's big enough (some topics will be almost unused)
- stopword list is also a hyperparameter

Outline.

LDA - the generative story, revisited

for topic $z \in \{1 \ldots |Z|\}$ **do**
  $\Phi_z \sim Dirichlet(\beta)$
end for
for document $d \in \{1 \ldots |D|\}$ **do**
  $\theta_d \sim Dirichlet(\alpha)$
  for word $i \in d$ **do**
    $z_i \sim Multinomial(\theta_d)$
    $w_i \sim Multinomial(\Phi_{z_i})$
  end for
end for

LDA - the maths, revisited

- The joint distribution of observed and hidden variables is:

$$P(D, \mathbf{z}, \Phi, \theta; \alpha, \beta) = \prod_{d \in D} p(\theta_d; \alpha) \prod_{i \in d} P(z_i; \theta_d) P(w_i; \Phi_z)$$

- and the posterior distribution is estimated via Gibbs sampling . . .

More on LDA and topic modelling

- Griffiths and Stevyers (2004): in the readings
- More diagrams/explanation:
  `http://psiexp.ss.uci.edu/research/papers/SteyversGriffithsLSABookFormatted.pdf`
- Diarmuid's slides:
  `https://www.cl.cam.ac.uk/~do242/Teaching/HIT-MSRA-2011/harbin_semantics_part2.pdf`
- MALLET: various software, including topic modelling
  `http://mallet.cs.umass.edu/topics.php` (with tutorial).

Next lectures

Rough plan:

- Restricted Bolzmann Machines
- RNNs and LSTMs
- distributonal models and word2vec