Introduction to topic models: Building up towards LDA

Vivi Nastase

Summer semester 2012

ICL, University of Heidelberg

Motivation

Assumption: the data was generated according to some model that we try to retrieve.

Motivation

Assumption: the data was generated according to some model that we try to retrieve.

The **observation** $\mathcal{O}$: the texts.
The **model** μ: parameters and probability distributions that generated the data – the one that fits the data best.

$$p(\mathcal{O}, \mu) \quad = \quad \underbrace{p(\mathcal{O}|\mu)}_{\textbf{likelihood}} \quad \underbrace{p(\mu)}_{\textbf{prior}} \quad = \quad \underbrace{p(\mu|\mathcal{O})}_{\textbf{posterior}} \quad \underbrace{p(\mathcal{O})}_{\textbf{evidence}}$$

Motivation

Assumption: the data was generated according to some model that we try to retrieve.

The **observation** $\mathcal{O}$: the texts.
The **model** μ: parameters and probability distributions that generated the data – the one that fits the data best.

$$p(\mathcal{O}, \mu) = \underbrace{p(\mathcal{O}|\mu)}_{\textbf{likelihood}} \underbrace{p(\mu)}_{\textbf{prior}} = \underbrace{p(\mu|\mathcal{O})}_{\textbf{posterior}} \underbrace{p(\mathcal{O})}_{\textbf{evidence}}$$

Best fit:

- maximize likelihood
- maximize the posterior
- compute the probability distribution of the posterior

Maximum likelihood estimation (ML)

Likelihood:

$$L(\mu|\mathcal{O}) = p(\mathcal{O}|\mu) = \prod_{x \in \mathcal{O}} p(x|\mu)$$

$$\mathcal{L}(\mu|\mathcal{O}) = \log L(\mu|\mathcal{O}) = \sum_{x \in \mathcal{O}} \log p(x|\mu)$$

Maximum likelihood estimation (ML)

Likelihood:

$$L(\mu|\mathcal{O}) = p(\mathcal{O}|\mu) = \prod_{x \in \mathcal{O}} p(x|\mu)$$

$$\mathcal{L}(\mu|\mathcal{O}) = \log L(\mu|\mathcal{O}) = \sum_{x \in \mathcal{O}} \log p(x|\mu)$$

The model that maximizes the likelihood:

$$\mu_{ML} = argmax_{\mu} \mathcal{L}(\mu|\mathcal{O}) = argmax_{\mu} \sum_{x \in \mathcal{O}} \log p(x|\mu)$$

Maximum likelihood estimation (ML)

Likelihood:

$$L(\mu|\mathcal{O}) = p(\mathcal{O}|\mu) = \prod_{x \in \mathcal{O}} p(x|\mu)$$

$$\mathcal{L}(\mu|\mathcal{O}) = \log L(\mu|\mathcal{O}) = \sum_{x \in \mathcal{O}} \log p(x|\mu)$$

The model that maximizes the likelihood:

$$\mu_{ML} = argmax_{\mu} \mathcal{L}(\mu|\mathcal{O}) = argmax_{\mu} \sum_{x \in \mathcal{O}} \log p(x|\mu)$$

Maximum:

$$\frac{\partial \mathcal{L}(\mu|\mathcal{O})}{\partial \mu_k} = 0; \forall \mu_k \in \mu$$

Maximum likelihood estimation (ML) – example

$\mathcal{O}$: N coin tossing experiments, $N = n_h + n_t$

Maximum likelihood estimation (ML) – example

$\mathcal{O}$: N coin tossing experiments, $N = n_h + n_t$

μ: p_h $\quad p(X = x_i | p_h) = p_h^{x_i}(1 - p_h)^{1-x_i}$

$x_i = 1$ for heads, $x_i = 0$ for tails

Maximum likelihood estimation (ML) – example

$\mathcal{O}$: N coin tossing experiments, $N = n_h + n_t$

μ: p_h $\quad p(X = x_i | p_h) = p_h^{x_i}(1 - p_h)^{1-x_i}$

$x_i = 1$ for heads, $x_i = 0$ for tails

$$\mathcal{L}(p_h|N) = \sum_{i=1,N} \log p(X = x_i|p_h)$$

Maximum likelihood estimation (ML) – example

$\mathcal{O}$: N coin tossing experiments, $N = n_h + n_t$

μ: p_h $\quad p(X = x_i | p_h) = p_h^{x_i}(1 - p_h)^{1-x_i}$

$x_i = 1$ for heads, $x_i = 0$ for tails

$$
\begin{aligned}
\mathcal{L}(p_h|N) &= \sum_{i=1,N} \log p(X = x_i|p_h) \\
&= \sum_{i=1,N} \log(p_h^{x_i}(1 - p_h)^{1-x_i}) \\
&= \sum_{i=1,N} x_i \log p_h + (1 - x_i)\log(1 - p_h) \\
&= n_h \log p_h + n_t \log(1 - p_h)
\end{aligned}
$$

$$
\begin{aligned}
p_{h\ ML} &= argmax_p \mathcal{L}(p_h|N) \\
&\rightarrow \frac{\partial \mathcal{L}(p_h|N)}{\partial p_h} = \frac{n_h}{p_{h\ ML}} - \frac{n_t}{1 - p_{h\ ML}} = 0
\end{aligned}
$$

Maximum a posteriori estimation (MAP)

Similar to ML estimation, but incorporates some prior knowledge about the parameters.

$$p(\mathcal{O}, \mu) = \underbrace{p(\mathcal{O}|\mu)}_{\textbf{likelihood}} \; \underbrace{p(\mu)}_{\textbf{prior}} = \underbrace{p(\mu|\mathcal{O})}_{\textbf{posterior}} \; \underbrace{p(\mathcal{O})}_{\textbf{evidence}}$$

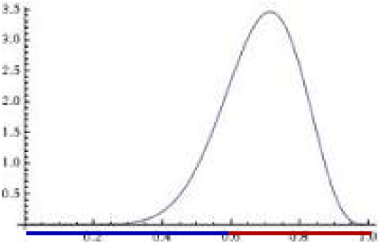
$$p(\mu|\mathcal{O}) = \frac{p(\mathcal{O}|\mu) \; p(\mu)}{p(\mathcal{O})}$$

Maximum a posteriori estimation (MAP)

Similar to ML estimation, but incorporates some prior knowledge about the parameters.

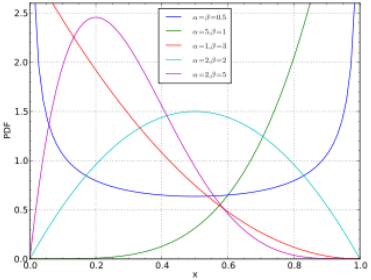
$$
\begin{aligned}
\mu_{MAP} &= argmax_{\mu} p(\mu|\mathcal{O}) \\
&= argmax_{\mu} \frac{p(\mathcal{O}|\mu)p(\mu)}{p(\mathcal{O})} \\
&= argmax_{\mu} \underbrace{p(\mathcal{O}|\mu)}_{\text{likelihood}} \underbrace{p(\mu)}_{\text{prior}} \\
&= argmax_{\mu} \mathcal{L}(\mathcal{O}|\mu) + \log p(\mu)
\end{aligned}
$$

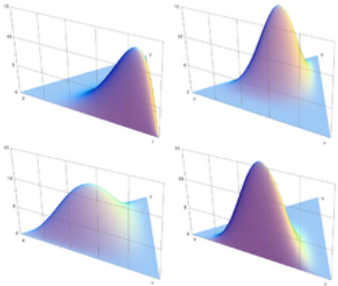
About priors – Beta/Dirichlet



$$p(p_h|\alpha_h, \alpha_t) = Beta(p_h|\alpha_h, \alpha_t) = \frac{1}{B(\alpha_h, \alpha_t)} p_h^{\alpha_h - 1}(1 - p_h)^{\alpha_t - 1}$$

About priors – Beta/Dirichlet





$$p(p_h|\alpha_h, \alpha_t) = Beta(p_h|\alpha_h, \alpha_t) = \frac{1}{B(\alpha_h, \alpha_t)} p_h^{\alpha_h - 1}(1 - p_h)^{\alpha_t - 1}$$

Maximum a posteriori estimation (MAP) – example

N coin tossing experiments, head/tails; $N = n_h + n_t$
p_h – probability to get a head
Prior belief about the coin: $p(p_h) = Beta(p_h|\alpha_h, \alpha_t)$

$$\begin{aligned}
\mu_{MAP} &= argmax_{\mu}\mathcal{L}(p_h|N) + \log p(p_h) \\
&\rightarrow \frac{n_h}{p_h} - \frac{n_t}{1-p_h} + \frac{\alpha_h - 1}{p_h} - \frac{\alpha_t - 1}{1-p_h} = 0 \\
\mu_{MAP} &= \frac{n_h + \alpha_h - 1}{N + \alpha_h - 1 + \alpha_t - 1}
\end{aligned}$$

Bayesian estimation

Similar to MAP estimation, but instead of maximizing to obtain the model, we induce a distribution of the model:

$$p(\mu|\mathcal{O}) = \frac{p(\mathcal{O}|\mu)p(\mu)}{p(\mathcal{O})}$$

The probability of the observation ($\mathcal{O}$) is the expected value, according to all possible model variations, given its prior:

$$p(\mathcal{O}) = \int_{\mu \in \theta} p(\mathcal{O}|\mu)p(\mu)d\mu$$

More about priors – conjugate distributions

$$p(\mu|\mathcal{O}) = \frac{p(\mathcal{O}|\mu)p(\mu)}{\int_{\mu\in\theta} p(\mathcal{O}|\mu)p(\mu)d\mu}$$

More about priors – conjugate distributions

$$p(\mu|\mathcal{O}) = \frac{p(\mathcal{O}|\mu)p(\mu)}{\int_{\mu\in\theta} p(\mathcal{O}|\mu)p(\mu)d\mu}$$

A conjugate prior $p(\mu)$ of a likelihood $p(\mathcal{O}|\mu)$ is a distribution that results in a posterior distribution $p(\mu|\mathcal{O})$ with the same functional form as the prior, and a parametrisation that incorporates the observations $\mathcal{O}$. (they are conjugate if they have the same functional form)

Discrete	Continuous
Bernoulli	Beta
$p(x) = q^x(1-q)^{1-x}$	$p(q) = \frac{1}{B(\alpha,\beta)}q^{\alpha-1}(1-q)^{\beta-1}$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 q^{\alpha-1}(1-q)^{\beta-1}$$

More about priors – conjugate distributions

$$p(\mu|\mathcal{O}) = \frac{p(\mathcal{O}|\mu)p(\mu)}{\int_{\mu\in\theta} p(\mathcal{O}|\mu)p(\mu)d\mu}$$

Discrete	Continuous
Bernoulli $p(x) = q^x(1-q)^{1-x}$	Beta $p(q) = \frac{1}{B(\alpha,\beta)} q^{\alpha-1}(1-q)^{\beta-1}$
Multinomial ...	Dirichlet ...
Poisson $p(x) = \frac{\lambda^x}{x!}e^{-\lambda}$	Gamma $p(\lambda) = \frac{1}{\theta^k}\frac{\lambda^{k-1}}{\Gamma(k)}e^{-\frac{\lambda}{\theta}}$

Probability of the observation – coin tossing

$$p(\mathcal{O}|\alpha_h, \alpha_t) = \int_0^1 p(\mathcal{O}|p_h)p(p_h|\alpha_h, \alpha_t)dp_h$$

Probability of the observation – coin tossing

$$\begin{aligned}
p(\mathcal{O}|\alpha_h, \alpha_t) &= \int_0^1 p(\mathcal{O}|p_h)p(p_h|\alpha_h, \alpha_t)dp_h \\
&= \int_0^1 p_h^{n_h}(1-p_h)^{n_t}\frac{1}{B(\alpha_h, \alpha_t)}p_h^{\alpha_h-1}(1-p_h)^{\alpha_t-1}dp_h \\
&= \frac{1}{B(\alpha_h, \alpha_t)}\int_0^1 p_h^{n_h+\alpha_h-1}(1-p_h)^{n_t+\alpha_t-1}dp_h \\
&= \frac{1}{B(\alpha_h, \alpha_t)}B(n_h+\alpha_h, n_t+\alpha_t)
\end{aligned}$$

Probability of the observation – coin tossing

$$
\begin{aligned}
p(\mathcal{O}|\alpha_h, \alpha_t) &= \int_0^1 p(\mathcal{O}|p_h) p(p_h|\alpha_h, \alpha_t) dp_h \\
&= \int_0^1 p_h^{n_h} (1 - p_h)^{1-n_t} \frac{1}{B(\alpha_h, \alpha_t)} p_h^{\alpha_h - 1} (1 - p_h)^{\alpha_t - 1} dp_h \\
&= \frac{1}{B(\alpha_h, \alpha_t)} \int_0^1 p_h^{n_h + \alpha_h - 1} (1 - p_h)^{n_t + \alpha_t - 1} dp_h \\
&= \frac{1}{B(\alpha_h, \alpha_t)} B(n_h + \alpha_h, n_t + \alpha_t)
\end{aligned}
$$

$$
B(\alpha_h, \alpha_t) = \frac{\Gamma(\alpha_h)\Gamma(\alpha_t)}{\Gamma(\alpha_h + \alpha_t)}
$$

$$
\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)
$$

Predicting a new observation

$$p(x = 1|\mathcal{O}, \alpha_h, \alpha_t) \quad = \quad \frac{p(x = 1, \mathcal{O}|\alpha_h, \alpha_t)}{p(\mathcal{O}|\alpha_h, \alpha_t)}$$

$$p(\mathcal{O}|\alpha_h, \alpha_t) = \frac{1}{B(\alpha_h, \alpha_t)} B(n_h + \alpha_h, n_t + \alpha_t)$$

Predicting a new observation

$$
\begin{aligned}
p(x = 1|\mathcal{O}, \alpha_h, \alpha_t) &= \frac{p(x = 1, \mathcal{O}|\alpha_h, \alpha_t)}{p(\mathcal{O}|\alpha_h, \alpha_t)} \\
&= \frac{\frac{1}{B(\alpha_h,\alpha_t)} B(n_h + 1 + \alpha_h, n_t + \alpha_t)}{\frac{1}{B(\alpha_h,\alpha_t)} B(n_h + \alpha_h, n_t + \alpha_t)}
\end{aligned}
$$

Predicting a new observation

$$
\begin{aligned}
p(x=1|\mathcal{O},\alpha_h,\alpha_t) &= \frac{p(x=1,\mathcal{O}|\alpha_h,\alpha_t)}{p(\mathcal{O}|\alpha_h,\alpha_t)} \\
&= \frac{\frac{1}{B(\alpha_h,\alpha_t)}B(n_h+1+\alpha_h,n_t+\alpha_t)}{\frac{1}{B(\alpha_h,\alpha_t)}B(n_h+\alpha_h,n_t+\alpha_t)} \\
&= \frac{B(n_h+1+\alpha_h,n_t+\alpha_t)}{B(n_h+\alpha_h,n_t+\alpha_t)} \\
&= \frac{\frac{\Gamma(n_h+1+\alpha_h)\Gamma(n_t+\alpha_t)}{\Gamma(N+1+\alpha_h+\alpha_t)}}{\frac{\Gamma(n_h+\alpha_h)\Gamma(n_t+\alpha_t)}{\Gamma(N+\alpha_h+\alpha_t)}}
\end{aligned}
$$

Predicting a new observation

$$
\begin{aligned}
p(x=1|\mathcal{O},\alpha_h,\alpha_t) &= \frac{p(x=1,\mathcal{O}|\alpha_h,\alpha_t)}{p(\mathcal{O}|\alpha_h,\alpha_t)} \\
&= \frac{\frac{1}{B(\alpha_h,\alpha_t)}B(n_h+1+\alpha_h,n_t+\alpha_t)}{\frac{1}{B(\alpha_h,\alpha_t)}B(n_h+\alpha_h,n_t+\alpha_t)} \\
&= \frac{B(n_h+1+\alpha_h,n_t+\alpha_t)}{B(n_h+\alpha_h,n_t+\alpha_t)} \\
&= \frac{\frac{\Gamma(n_h+1+\alpha_h)\Gamma(n_t+\alpha_t)}{\Gamma(N+1+\alpha_h+\alpha_t)}}{\frac{\Gamma(n_h+\alpha_h)\Gamma(n_t+\alpha_t)}{\Gamma(N+\alpha_h+\alpha_t)}} \\
&= \frac{n_h+\alpha_h}{N+\alpha_h+\alpha_t}
\end{aligned}
$$

$$\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$$

Bayesian estimation – example

$\mathcal{O}$: A sequence of N coin tosses, $N = n_h + n_t$

μ: $p_h | \alpha_h, \alpha_t$

$$\begin{aligned} p(\mu|\mathcal{O}) &= \frac{p(\mathcal{O}|\mu)p(\mu)}{p(\mathcal{O})} \\ &= \frac{\prod_{i=1,N} p(X = x_i|p_h)p(p_h|\alpha_h, \alpha_t)}{\int_0^1 \prod_{i=1,N} p(X = x_i|p_h)p(p_h|\alpha_h, \alpha_t)dp_h} \end{aligned}$$

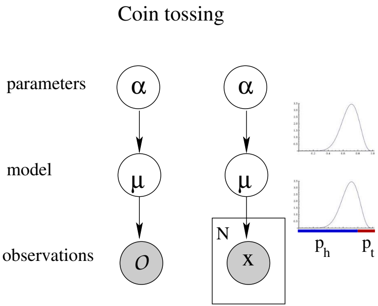
Bayesian estimation – example

$\mathcal{O}$: A sequence of N coin tosses, $N = n_h + n_t$

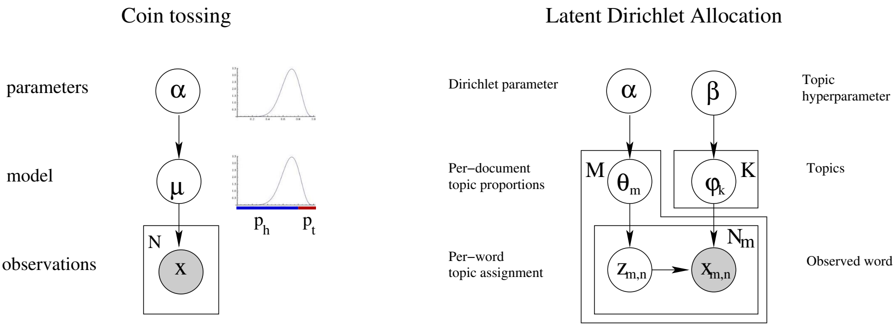
μ: $p_h|\alpha_h, \alpha_t$

$$
\begin{aligned}
p(\mu|\mathcal{O}) &= \frac{p(\mathcal{O}|\mu)p(\mu)}{p(\mathcal{O})} \\
&= \frac{\prod_{i=1,N} p(X = x_i|p_h)p(p_h|\alpha_h, \alpha_t)}{\int_0^1 \prod_{i=1,N} p(X = x_i|p_h)p(p_h|\alpha_h, \alpha_t)dp_h} \\
&= \frac{p_h^{n_h}(1-p_h)^{n_t} \quad \frac{1}{B(\alpha_h,\alpha_t)} p_h^{\alpha_h - 1}(1-p_h)^{\alpha_t - 1}}{\frac{1}{B(\alpha_h,\alpha_t)} B(n_h + \alpha_h, n_t + \alpha_t)} \\
&= \frac{1}{B(n_h + \alpha_h, n_t + \alpha_t)} p_h^{n_h + \alpha_h - 1}(1-p_h)^{n_t + \alpha_t - 1} \\
&= Beta(p_h|n_h + \alpha_h, n_t + \alpha_t)
\end{aligned}
$$

Bayesian network

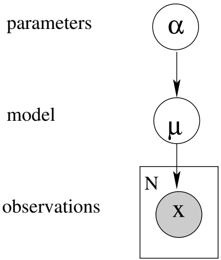


Latent Dirichlet Allocation

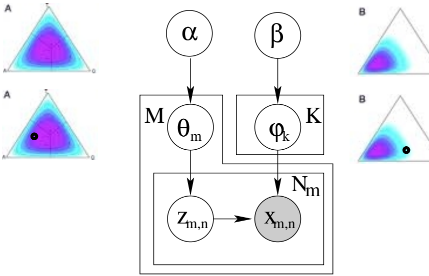


Latent Dirichlet Allocation

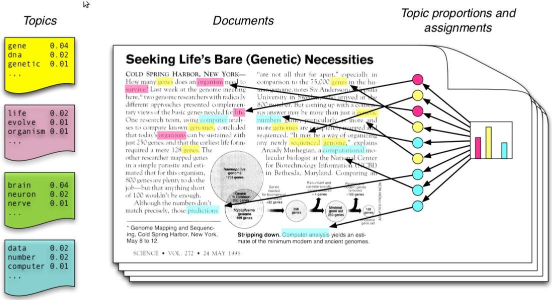
Coin tossing



Latent Dirichlet Allocation



Topic models



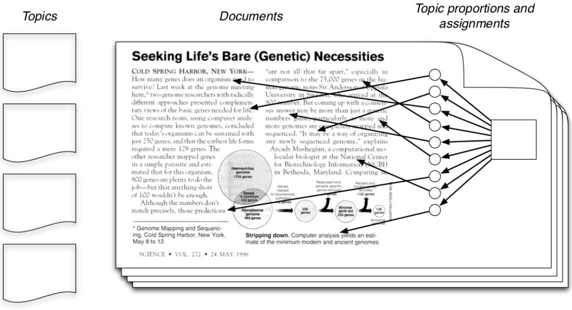
Each document is a mixture of topics:

$$\sum_k p(z_m = k) = \sum_k \theta_{m,k} = 1$$

Each word is drawn from one of its document's topics:

$$p(w_{m,n}) = \sum_k p(w_{m,n} | z_{m,n} = k) p(z_{m,n} = k) = \sum_k \phi_k(w_{m,n}) \theta_{m,k}$$

Topic models



The **observations** are the documents: $\mathbf{w_m}, m \in 1, M$
We need to infer the **model**, i.e the underlying topic structure, i.e. the topic assignments $z_{m,n}$, the topic $\theta_m, \quad m \in 1, M$ and word distributions $\phi_k, \quad k \in 1, K$

Priors:

$\theta \sim$ distribution with hyperparameter α

$\phi \sim$ distribution with hyperparameter β

Topic models – Latent Dirichlet Allocation

$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_k \theta_k^{\alpha_k - 1}$$

$$\sum_k \theta_{m,k} = 1$$

α controls the mean shape and sparsity of θ

The topic proportions (θ_m) are a K-dimensional Dirichlet

$z_{m,n}$ are multinomial distributions from θ_m

$$p(z_{m,n}|\theta_m) = \frac{N!}{\prod_{k=1}^{K} n_k!} \prod_{k=1}^{K} \theta_{m,k}^{n_k}$$

Topic models – Latent Dirichlet Allocation

$$p(\phi|\beta) = \frac{1}{B(\beta)} \prod_v \phi_v^{\beta_v - 1}$$

$$\sum_v \phi_{k,v} = 1$$

β controls the mean shape and sparsity of ϕ
The topics (ϕ_k) are a V-dimensional Dirichlet
$w_{m,n}$ are multinomial distributions from $\phi_{z_{m,n}}$

$$p(w_{m,n}|\phi_k) = \frac{V!}{\prod_{v=1}^{V} n_v!} \prod_{v=1}^{V} \phi_{k,v}^{n_v}$$

Topic models

Remember: $p(x=1|\mathcal{O},\alpha_h,\alpha_t) = \frac{p(x=1,\mathcal{O}|\alpha_h,\alpha_t)}{p(\mathcal{O}|\alpha_h,\alpha_t)} = \frac{n_h+\alpha_h}{N+\alpha_h+\alpha_t}$

$$p(z_i = k, w_i|\mathbf{z_{-i}}) = p(z_i = k|\mathbf{z_{-i}})p(w_i|z_i)$$

$$\begin{aligned} p(z_i = k|\mathbf{z_{-i}},\alpha) &= p(z_i = k|\mathbf{z_{-i}},\alpha) \\ &= \frac{nz^k_{m,-i} + \alpha_k}{\sum_{j=1}^{K}(nz^j_{m,-i} + \alpha_j)} \\ &= \hat{\theta}_{m,k} \quad \text{an approximation of } \theta_{m,k} \end{aligned}$$

$$\begin{aligned} p(w_i|z_i,\beta) &= \frac{nw^{z_i}_{w_i,-i} + \beta_{w_i}}{\sum_{l=1}^{V}(nw^{z_i}_l + \beta_l)} \\ &= \hat{\phi}_k(w_i) \quad \text{an approximation of } \phi_k(w_i) \end{aligned}$$

Sample z_i and assign it at possition i: $z_i \propto \hat{\theta}_{m,k}\hat{\phi}_k(w_i)$

Inference via Gibbs sampling

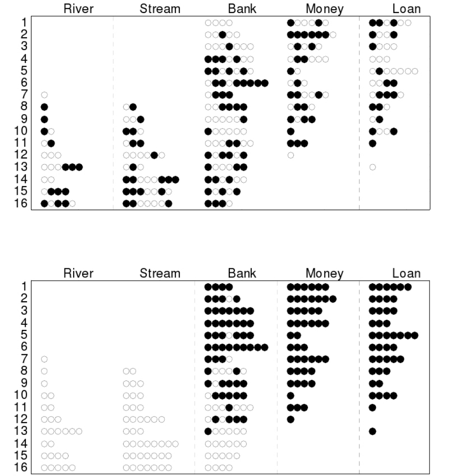


Figure 7. An example of the Gibbs sampling procedure.

Sample z_i and assign it at possition i: $z_i \propto \hat{\theta}_{m,k}\hat{\phi}_k(w_i)$

References

- *Probabilistic topic models*, Mark Steyvers, Tom Griffiths
- *Parameter estimation for text analysis*, Gregor Heinrich
- *Topic Models*, David Blei (tutorial, videolectures.net)