Introducing Bayesian Language models

Unigram, Bigram, Hidden Markov, & Topic models

Mark Andrews
Psychology Department, Nottingham Trent University

✉ mark.andrews@ntu.ac.uk
🐦 @xmjandrews

July 16, 2017

Language data

▶ A *minimal* description of observed language data is as follows:

$$\mathcal{D} = \{w_1, w_2 \ldots w_j \ldots w_J\},$$

with

$$w_j = w_{ji}, w_{j2} \ldots w_{jn_j},$$

and

$$w_{ji} \in \mathcal{V},$$

where $\mathcal{V}$ is a finite vocabulary, which can be represented for simplicity by integers

$$\mathcal{V} = \{1, 2 \ldots V\}.$$

▶ In other words, we can describe language data as a set of sequences of symbols.

▶ Each w_{ji} usually represents a word (but could represent e.g. phonemes, instead) and each w_j represents a text or sentence.

Unigram probabilistic language model

- A *probabilistic language model* is any probabilistic generative model of the language data $\mathcal{D}$.
- One of the simplest possible probabilistic language models is

$$w_{ji} \sim \text{dcat}(\pi), \quad \text{for } i \in 1, 2 \ldots n_j, j \in 1, 2 \ldots J$$

where dcat( ) represents a categorical distribution (probability mass function) over $1 \ldots V$ with

$$\Pr(w_{ji} = k) = \pi_k.$$

- In other words, we model each word w_{ji} as drawn independently from a single categorical distribution with parameters π.

Bayesian unigram models

- In any Bayesian model, we provide a full probabilistic account of all variables, including observed variables, parameters, etc.
- For the case of a probabilistic unigram model, a common Bayesian unigram model would be

$$\pi \sim \text{ddirichlet}(\alpha),$$
$$w_{ji} \sim \text{dcat}(\pi), \quad \text{for } i \in 1, 2 \dots n_j, j \in 1, 2 \dots J.$$

- Here, ddirichlet( ) denotes a V-dimensional Dirichlet distribution, which is a probability distribution over V-dimensional probability mass functions. It has hyperparameters $\alpha = \alpha_1, \alpha_2 \dots \alpha_v \dots \alpha_V$, where each $\alpha_v > 0.0$.
- The Dirichlet distribution is commonly chosen in this context because it is a conjugate prior of the categorical distribution.

Dirichlet Distribution

- The Dirichlet distribution is

$$\text{ddirichlet}(\alpha) = \frac{\Gamma(\sum_v \alpha_v)}{\prod_v \Gamma(\alpha_v)} \prod_{v=1}^{V} \pi_v^{\alpha_v - 1}$$

where

$$\frac{\Gamma(\sum_v \alpha_v)}{\prod_v \Gamma(\alpha_v)}$$

is the normalizing constant, i.e.,

$$\frac{\prod_v \Gamma(\alpha_v)}{\Gamma(\sum_v \alpha_v)} = \int \prod_{v=1}^{V} \pi_v^{\alpha_v - 1} d\pi$$
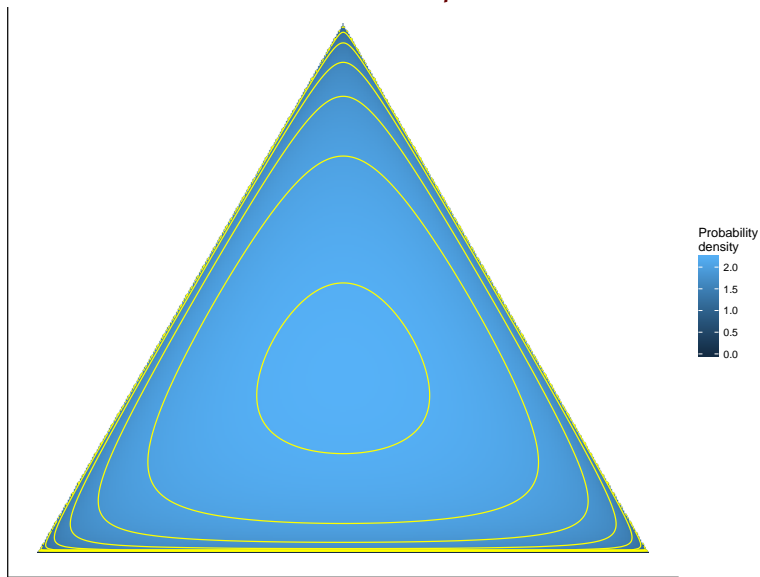
Dirichlet distribution hyperparameters

- As mentioned, the hyperparameters of a V-dimensional Dirichlet distribution is a V-dimensional vector α of positive values.
- It is convenient to sometimes represent α as follows:

$$\alpha = a \cdot m$$

  where $a > 0$ is a scalar and m is V-dimensional probability mass function, i.e $0 \leqslant m_v \leqslant 1$ and $\sum_v m_v = 1$.
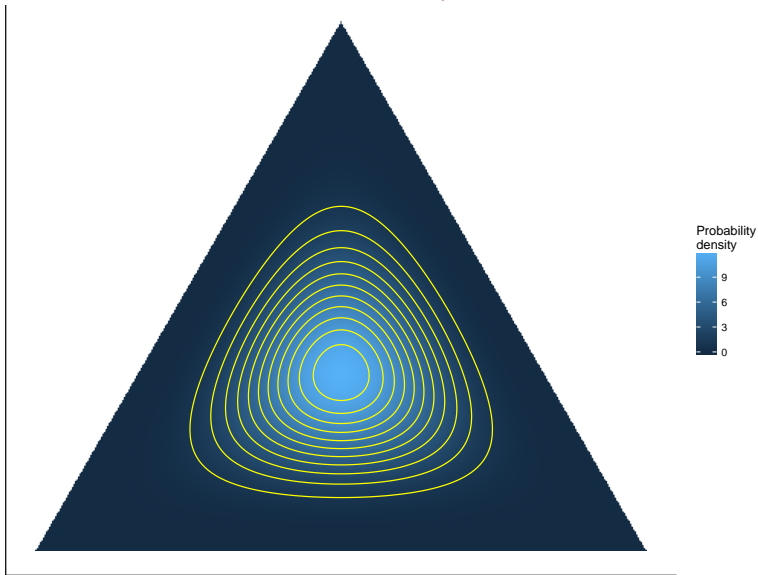- In this reparameterization, m is a *location* vector and a is a *scale* vector. In other words, m is the centre of the distribution, and a indicates how spread out around the centre the distribution is. Also, m is exactly the average, or expected value of the distribution.
- Note $a = \sum_v \alpha_v$, $m = \alpha/a$.

3d Dirichlet distribution: Example 1

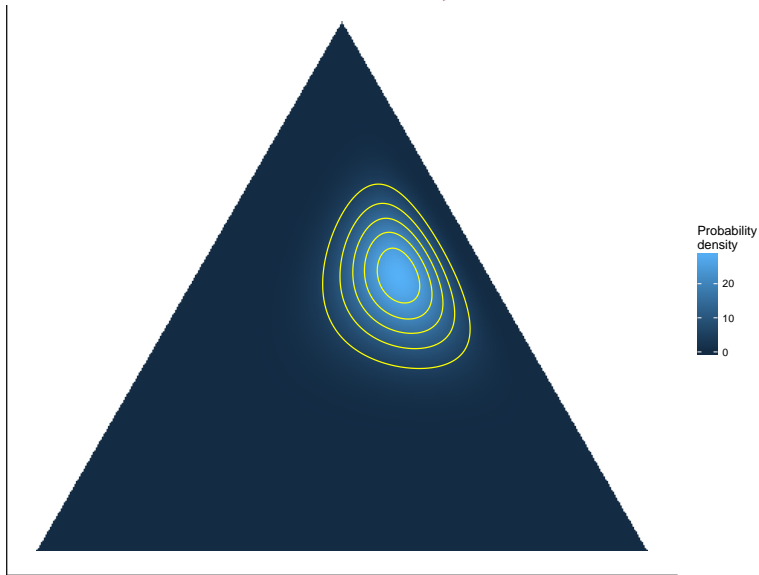Here, $\alpha = 1.1, 1.1, 1.1$, or $a = 3.3$, $m = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$

3d Dirichlet distribution: Example 2

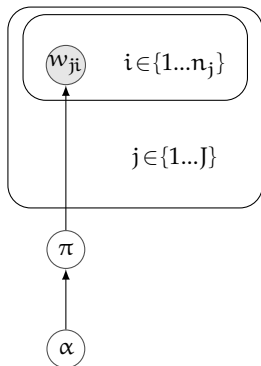Here, $\alpha = 5, 5, 5$, or $a = 15$, $m = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$

3d Dirichlet distribution: Example 3

Here, $\alpha = 5, 10, 15$, or $a = 30$, $m = [\frac{1}{6}, \frac{1}{3}, \frac{1}{2}]$

Unigram model: Bayesian network

This is a *graphical model* or *Bayesian network*. It shows the conditional independence structure of the variables in the probabilistic model.

Posterior inference

▶ Because the Dirichlet distribution is a conjugate prior to the categorical distribution, calculating the posterior distribution over π is possible algebraically:

$$P(\pi|\mathcal{D}, \alpha) = \frac{P(D|\pi)P(\pi|\alpha)}{\int P(D|\pi)P(\pi|\alpha)\,d\pi}, \propto \prod_{\{ji\}} P(w_{ji}|\pi)P(\pi|\alpha),$$

$$\propto \prod_{v=1}^{V} \pi_v^{n_v} \prod_{v=1}^{V} \pi_v^{\alpha_v - 1},$$

$$\propto \prod_{v=1}^{V} \pi_v^{n_v + \alpha_v - 1},$$

$$\mathrm{ddirichlet}(n + \alpha).$$

▶ In other words, given the prior $\mathrm{ddirichlet}(\alpha)$ and observed counts of $n = n_1, n_2 \ldots n_v$ (i.e., n_v is the number of observations of word v), the posterior distribution is always $\mathrm{ddirichlet}(n + \alpha)$.

Unigram model of Moby Dick

```r
data("stop_words")
moby_dick <- gutenberg_download(2701)


word_counts <- moby_dick %>%
  unnest_tokens(word, text, token = 'words') %>%
  anti_join(stop_words) %>%
  count(word, sort=TRUE)

# Make a (named) vector of counts
counts <- word_counts$n
words <- word_counts$word
```

Unigram model of Moby Dick

- As a starting point, choose uniform prior
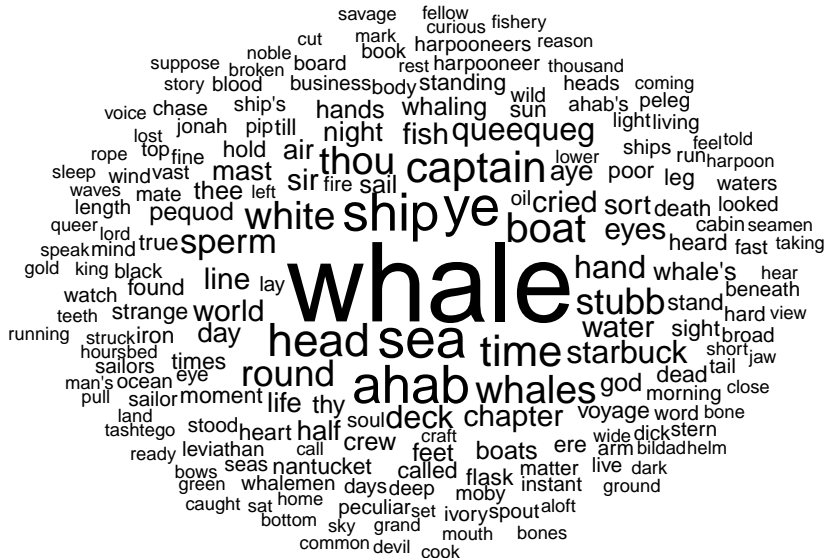
```
alpha <- rep(1, length(counts))
```

- Draw a sample from posterior distribution $P(\pi|\mathcal{D}, \alpha)$:
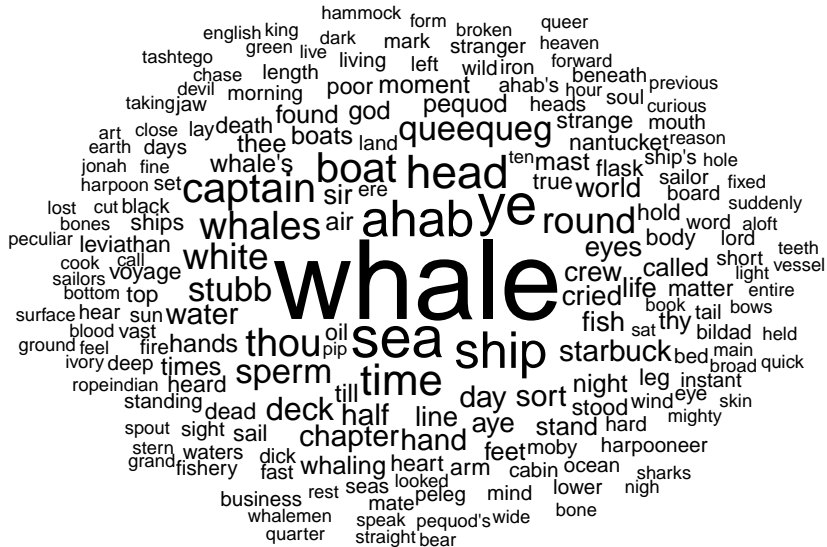
```
vpi <- sample.dirichlet(counts+alpha)
```

- We can graphically represent this sample as follows

```
wordcloud(words = words, freq = vpi, max.words = 200)
```
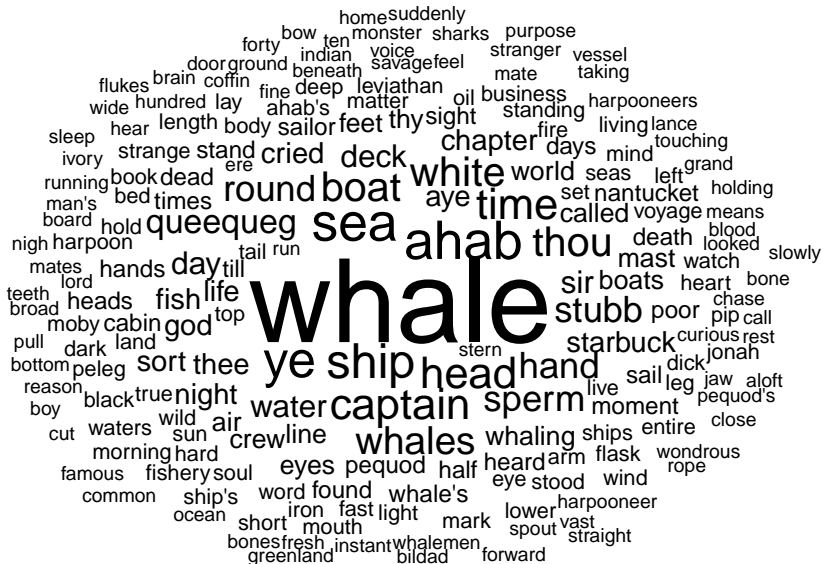
Sample from posterior distribution: Example 1

Sample from posterior distribution: Example 2

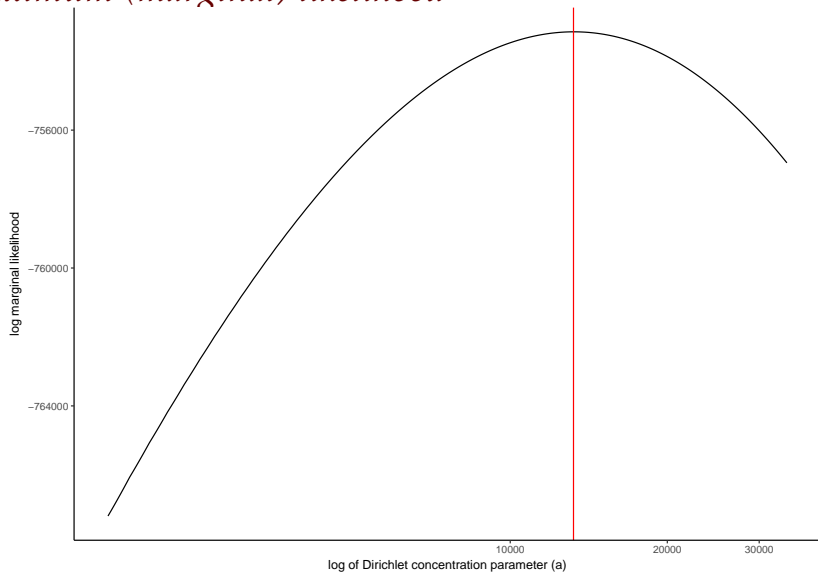*Sample from posterior distribution: Example 3*

Marginal likelihood

- The marginal likelihood is

$$P(\mathcal{D}|\alpha) = \int P(D|\pi)P(\pi|\alpha)\,d\pi,$$

$$= \frac{\Gamma(\sum_v \alpha_v)}{\prod_v \Gamma(\alpha_v)} \int \prod_{v=1}^{V} \pi_v^{n_k + \alpha_k - 1}\,d\pi,$$

$$= \frac{\Gamma(\sum_v \alpha_v)}{\prod_v \Gamma(\alpha_v)} \frac{\prod_v \Gamma(n_v + \alpha_v)}{\Gamma(\sum_v (\alpha_v + n_v))}.$$

- It is the likelihood function of α, and is marginalized (i.e. integrated) over π.
- We can treat this like any other likelihood function.

Maximum (marginal) likelihood

We can set the value of a to the value that maximizes the likelihood function. This is a type of *empirical Bayes* inference.

Bigram models

- A bigram probabilistic language model for our observed language data $\mathcal{D}$ is as follows.
- For $1 \leqslant j \leqslant J$,

$$w_{j1} \sim \text{dcat}(\pi),$$
$$w_{ji} \sim \text{dcat}(\theta_{[w_{ji-1}]}), \quad \text{for } 2 \leqslant i \leqslant n_j.$$

  where θ is a $V \times V$ matrix, and $\theta_{vu} \triangleq P(w_{ji} = u | w_{ji-1} = v)$

- In other words, for each sentence or text, we first sample the initial word for a categorical distribution. For the remaining words, we sample their values from a categorical distribution that is conditioned on the previous word.

Bayesian Bigram model

- For a Bayesian version of a bigram model, we provide priors on π and θ.
- Common choices are Dirichlet distributions

$$\pi \sim \text{ddirichlet}(\alpha_0),$$
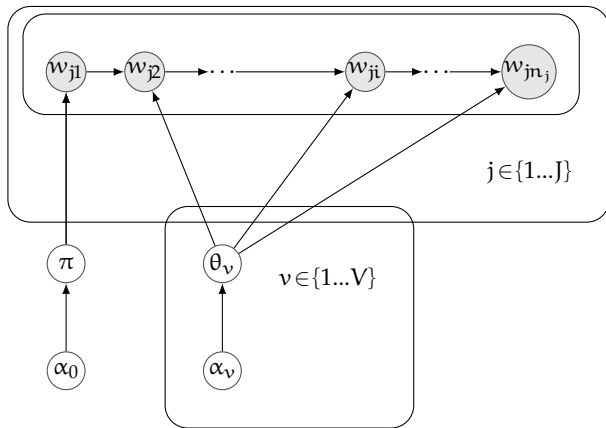$$\theta_v \sim \text{ddirichlet}(\alpha_v), \quad \text{for } 1 \leqslant v \leqslant V,$$
$$w_{j1} \sim \text{dcat}(\pi),$$
$$w_{ji} \sim \text{dcat}(\theta_{w_{ji-1}}), \quad \text{for } 2 \leqslant i \leqslant n_j.$$

- Note that here we have a separate α_v prior for each conditional distribution θ_v. Other options are possible.

Bigram model Bayesian network

Posterior inference

- The posterior distribution over θ_v is similar to the case of the posterior inference in Dirichlet distributions in the unigram case, with the difference being that the relevant *counts* are

$$R_v = R_{v1}, R_{v2} \ldots R_{vu} \ldots R_{vV}$$

  where R_{vu} gives the number of times word u follows word v in the corpus.

- As such,

$$P(\theta_v | \mathcal{D}) = \text{ddirichlet}(R_v + \alpha_v)$$

- For the posterior over π, we can use the marginal count vector r, where r_v simply gives the frequency of occurrence of word v:

$$P(\pi | \mathcal{D}) = \text{ddirichlet}(r + \alpha_0)$$

Bigram model of collected works of Charles Dickens

```r
dickens_books_id <- gutenberg_works(
  author == "Dickens, Charles")$gutenberg_id
dickens_books <- gutenberg_download(
  dickens_books_id, meta_fields = "title")

dickens_words <- dickens_books %>%
  unnest_tokens(word, text, token = 'words') %>%
  count(word, sort=TRUE)

dickens_bigrams <- dickens_books %>%
  unnest_tokens(bigram, text, token = 'ngrams', n=2) %>%
  count(bigram, sort=TRUE) %>%
  separate(bigram, c('word1', 'word2'), sep=' ')
```

Sample from posterior distribution of θ_{dear}

Sample from posterior distribution of θ_{stand}

Sample from posterior distribution of θ_{tall}

Simulations from a bigram model

- We can generate data from a trained bigram model as follows:
- First, we sample from the posterior distribution $P(\pi|\mathcal{D}, \alpha_0)$, and then sample w_1 from $\text{dcat}(\pi)$.
- We then sample from $P(\theta_{[w_1]}|\mathcal{D}, \alpha_{[w_1]})$, and sample w_2 from $\text{dcat}(\theta_{[w_1]})$.

Simulations from a bigram model: Examples 1-3

- ► admirable lady acknowledged supremacy and then i might think that ought to whom she once not your time organist ought to divine who had folded up to mutter among them a level of the next morning three j linton bespeckled prospect the knocker marley's slushed and defarge as mr jaggers
- ► curtains wherever he took deadly cold arch wore itself but i would have we my own face was quite necessary to wear this couplet about two pair of saying that in his vigorous play with an arm round upon that the real military life may be an eye these walls
- ► constantinopolitan chamber straight before they looked at enormous hit with me steerforth as they called him with the people live in it to night if he had entered according to do sit thee get'st tobacco post in germany and of characteristics of beasts and slept and no idea good part

Simulations from a bigram model: Examples 4-6

- only ended the duke has lost to this little tavern he has been careful of the door and present seat mr pickwick after making expiation for this three quarters over an angry sultan pomp about the plan of all appearance of yesterday i shut lest he covered with love the
- name as if they came to contribute to his gloved hand would be such a positive intelligence when he had insisted that he did not quite a wry faces clustering pillars strong alliance but she could not that it formed into the other side with me we find yourself to
- most distant corner of oh cap'en cuttle walked forth against me but whom he opened from view to him what provision to cast off a little cage in every fresh discussion fledgeby you do it the expression of the house where he observed to giving me and with this woman

Simulations from a bigram model: Examples 7-9

- ▶ boot cleaning for i would have so i had not set of course though it had come upon her and all very near your heart which i don't you should disparage my dear rogues you to a bar and full sincerity he much tickled in the mother and more money
- ▶ restless too much satisfaction of breath folding up directly in one of the obstinacy reputation and of us for i may derive from the third in all the regular in the park the top of the correction of that the change in reduced and find him and believe i should
- ▶ penalty then drank bitter laughs i need keep up into the confederation observes when i cannot come at the old curiosity out in reference to come to appear asleep what's her rosy cheeked charmers with great that the seat till now don't know that had said she gave some horrible

Hidden Markov model

- A Hidden Markov model is a type of latent variable model.
- It assumes that observed data are generated by sampling from categorical distributions that are indexed by a latent (hidden) first order Markov chain of state variables.
- The cardinality of the state variable, denoted by K here, is a modelling choice.
- As a probabilistic language model for our $\mathcal{D}$, a Hidden Markov model is a follows: For $j \in \{1 \dots J\}$,

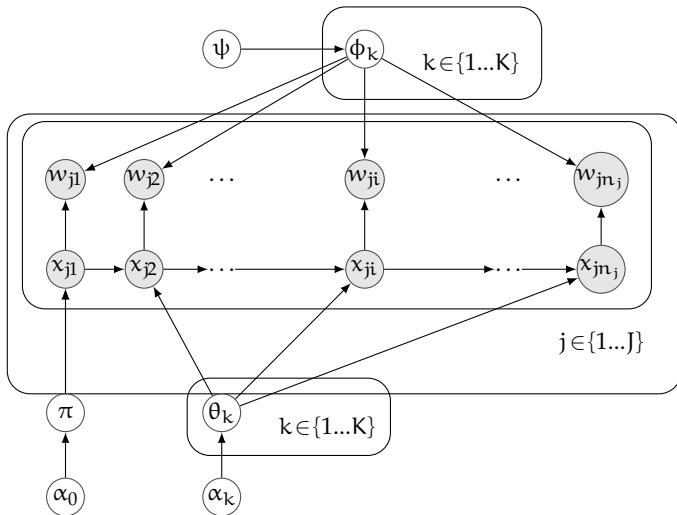$$x_{j1} \sim \text{dcat}(\pi),$$
$$x_{ji} \sim \text{dcat}(\theta_{[x_{ji-1}]}), \quad \text{for } 2 \leqslant i \leqslant n_{ji},$$
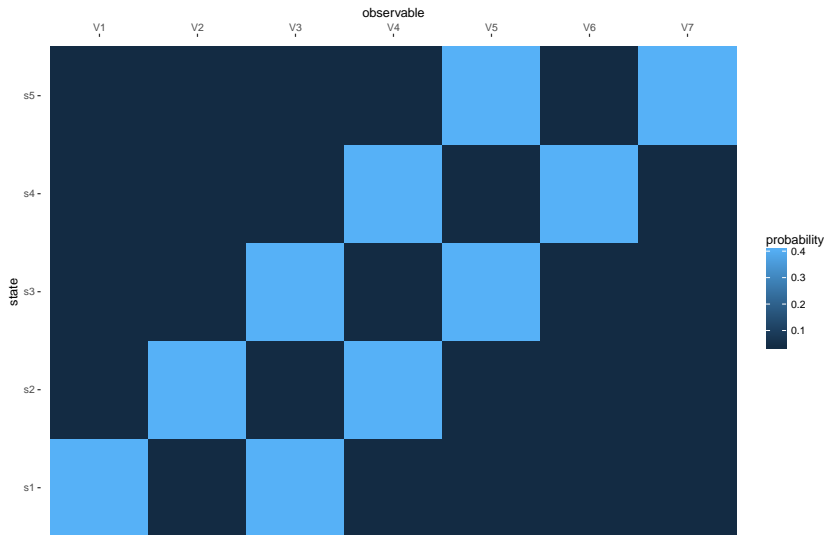$$w_{ji} \sim \text{dcat}(\phi_{[x_{ji}]}), \text{for } 2 \leqslant i \leqslant n_{ji}$$

- Here, θ is a $K \times K$ matrix, with $P(x_{ji} = l | x_{ji-1} = k) = \theta_{kl}$.
- The ϕ is a $K \times V$ matrix, $P(w_{ji} = v | x_{ji} = k) = \phi_{kv}$.
- We put Dirichlet priors on π, θ, ϕ.
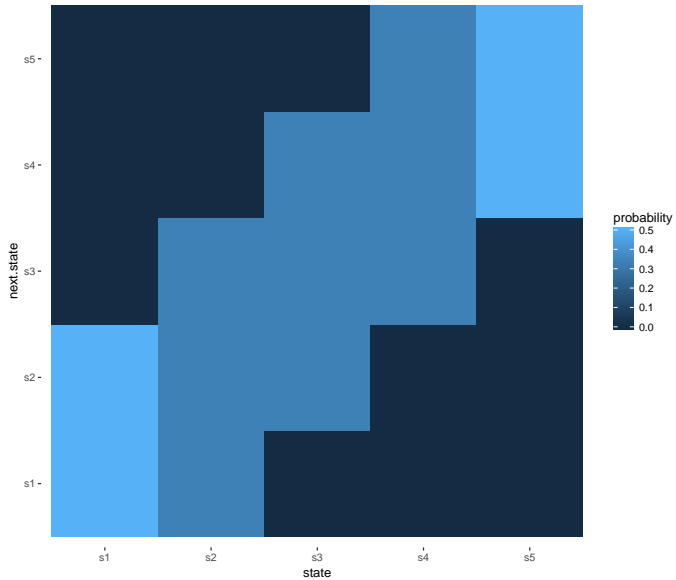
Hidden Markov model Bayesian network

Toy example: φ

Toy example: θ

Posterior inference

- ► We usually will use Dirichlet priors on π, $\phi_1, \phi_2 \ldots \phi_K$, and $\theta_1, \theta_2 \ldots \theta_K$.
- ► However, the posterior distribution

$$P(\pi, \phi, \theta | \mathcal{D})$$

  is not analytically tractable.

- ► In this situation, we use Monte Carlo methods to draw samples from this posterior distribution.
- ► For this, we may use a *blocked Gibbs sampler* such where we iteratively draw a sample from the posterior distribution of the latent state space, assuming values for the parameters, and then from the posterior over the parameters, assuming values for the state space:

$$\tilde{x} \sim P(x | w, \tilde{\pi}, \tilde{\phi}, \tilde{\theta}),$$
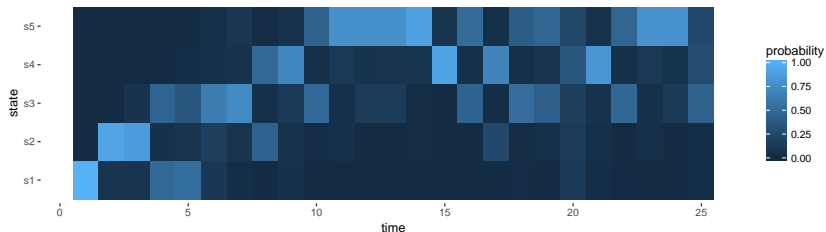$$\tilde{\pi}, \tilde{\phi}, \tilde{\theta} \sim P(\pi, \phi, \theta | w, \tilde{x})$$

State space inference

- Given a sequence of observations $w_1, w_2 \ldots w_n$, what is the probability distribution over the possible values of $x_1, x_2 \ldots x_n$?
- Assuming we know π, ϕ, θ, we can use recursive inference as follows:

$$P(x_i|w_1 \ldots w_2) = \frac{P(w_i|x_i)P(x_i|w_1, w_2 \ldots w_{i-1})}{\sum_{\{x_i\}} P(w_i|x_i)P(x_i|w_1, w_2 \ldots w_{i-1})},$$
$$\propto P(w_i|x_i) \sum_{\{x_{i-1}\}} P(x_i|x_{i-1})P(x_{i-1}|w_1, w_2 \ldots w_{i-1})$$

State space inference in toy example

▶ Given the observed values — 4, 2, 4, 3, 3, 5, 5, 4, 6, 5, 7, 5, 5, 7, 6, 5, 4, 5, 5, 1, 6, 5, 7, 5, 3 — the inferred values of the states are:

▶ The actual values of the state trajectory are: 1, 2, 3, 3, 3, 4, 5, 4, 4, 5, 5, 5, 5, 5, 4, 3, 2, 3, 3, 3, 4, 3, 2, 3, 3.

Latent Dirichlet Allocation

- Latent Dirichlet Allocation is a probabilistic bag-of-words language model.
- It is a type of multilevel/hierarchical probabilistic mixture model.
- It is defined as follows: For $1 \leqslant j \leqslant J$,

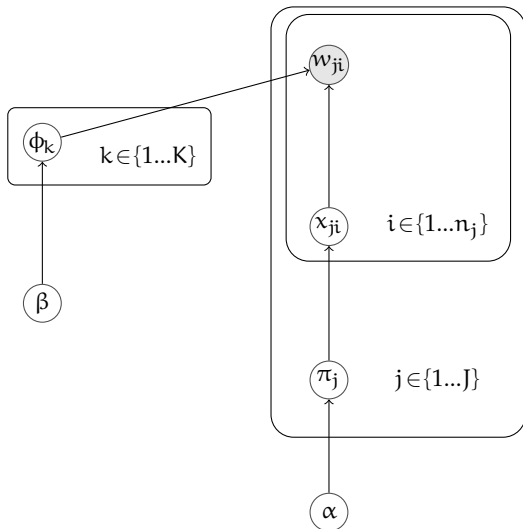$$\pi_j \sim \text{ddirichlet}(\alpha), \quad \text{for } 1 \leqslant j \leqslant J,$$

and for $1 \leqslant i \leqslant n_{ji}$,

$$x_{ji} \sim \text{dcat}(\pi_j),$$
$$w_{ji} \sim \text{dcat}(\phi_{[x_{ji}]})$$

- The cardinality of ϕ is a modelling choice.
- As such, we treat each document j as a sample from a probabilistic mixture model, where the component probabilities are π_j, which are sampled from a Dirichlet distribution.

Latent Dirichlet Allocation Bayesian network

- ▶ The Bayesian network diagram of the Latent Dirichlet Allocation model

Topic modelling the AP corpus

```r
data("AssociatedPress")

ap_lda <- LDA(AssociatedPress,
              k = 50,
              control = list(seed = 1234))

tidy(ap_lda, matrix = "beta") %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta) %>%
```
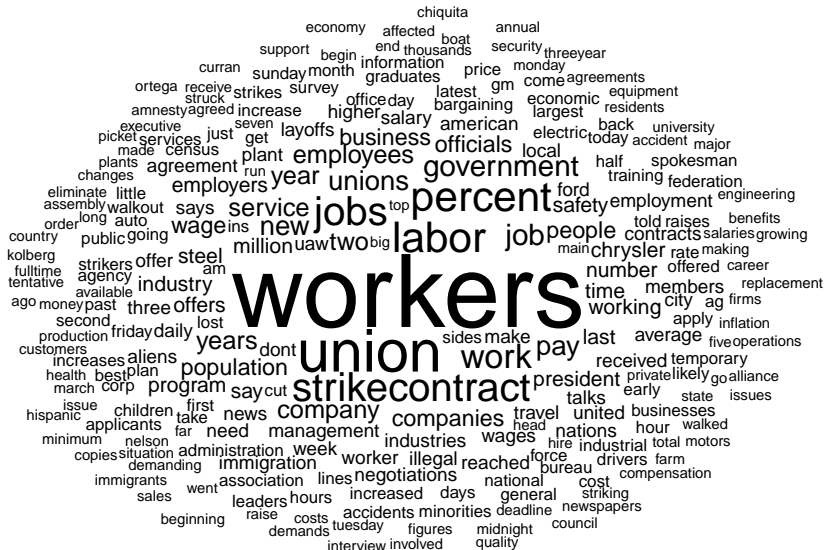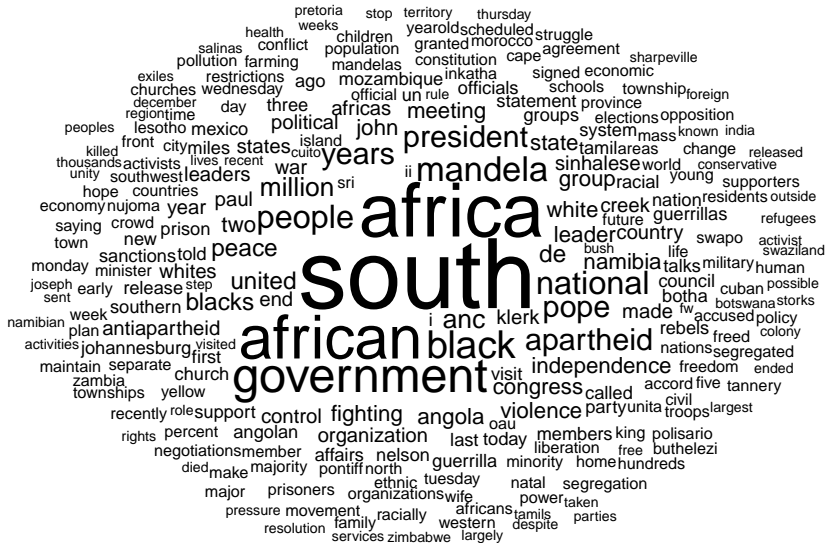
Example topics from AP corpus

topic_01	topic_05	topic_10	topic_15	topic_20
workers	south	i	party	drug
union	africa	president	government	i
labor	african	united	political	cocaine
percent	government	states	communist	police
jobs	black	people	soviet	case
strike	mandela	time	opposition	attorney
contract	people	vietnam	gorbachev	fbi
work	national	years	minister	court
job	president	support	elections	told
government	apartheid	government	leader	two

Topic 1

workers union labor jobs percent strike contract company strikers work pay president companies industry employees government service wage unions million labor job people year new employers steel population management immigration worker illegal reached negotiations national average members working time chrysler nations businesses wages drivers motors farm leaders hours increased days general minorities newspapers council midnight quality
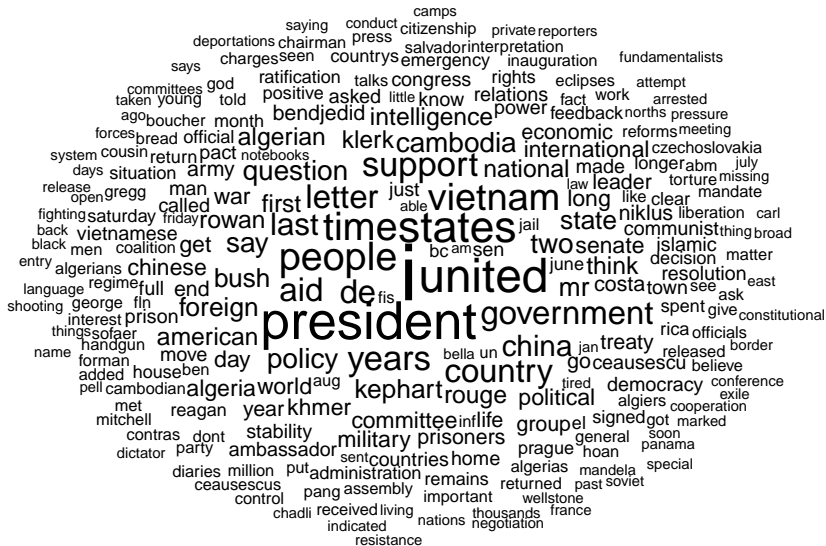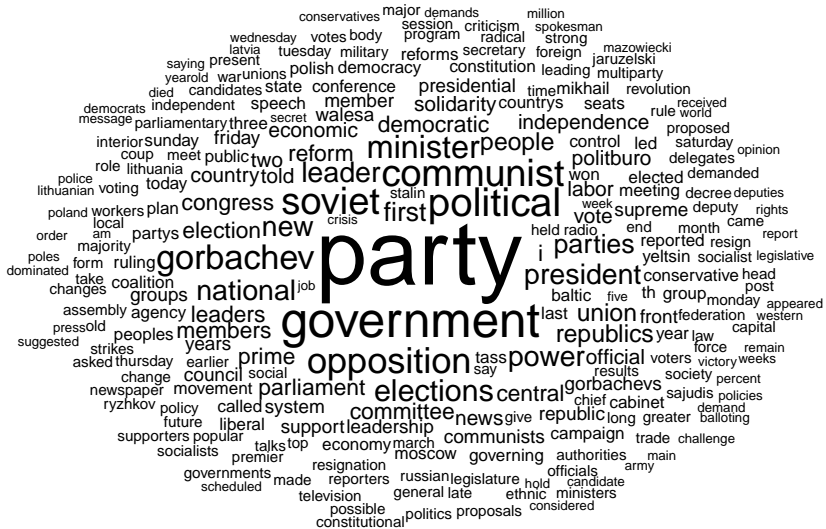
Topic 5

Topic 10

Topic 15

Topic 20