LINKÖPING UNIVERSITY

Latent Dirichlet Allocation

Johan Alenlöv, Linköping University

2023-10-31

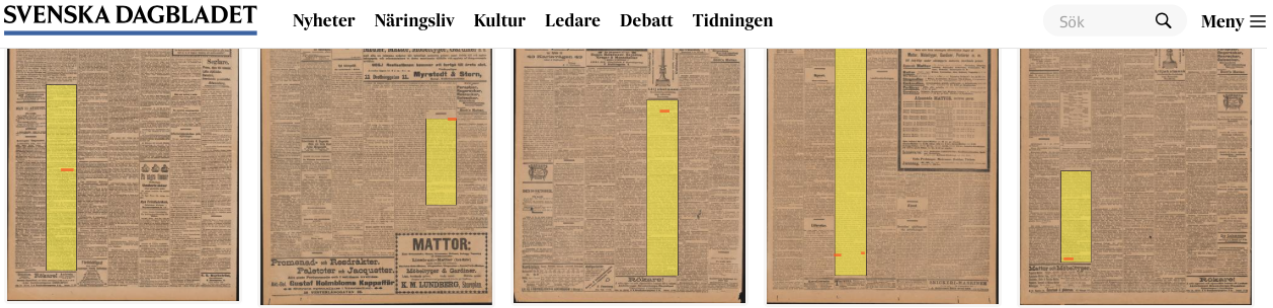
Outline – Lecture 5

Aim: Introduce the Latent Dirichlet Allocation model (LDA) and show how to perform inference in this model through collapsed Gibbs sampling.

Outline:

1. Topic modelling
2. The graphical model and plate notation.
3. Exploring a corpus using the LDA model.
4. Gibbs sampling in LDA.

Topic modelling



"...generalmajor Gadd den 24 dennes kL 1 e m Det litterära **Nobelpriset** Med anledning af meddelandet i en tysk..."

"...sålunda skulle röjt dem Svensk-amerikansk aspirant till **Nobelpriset** Hemlandets i Chikago redaktion..."

"...öfverläggningar om **Nobelpriset** Skog smedels fonden och järnvägsbyggandet K m :t har medgifvit att Orsa..."

"...Samarows längesedan vissnade lagrar icke ha lämnat honom någon ro Eller är det månne **Nobelpriset**..."

"...Svenska akademien sammanträder i morgon hvarvid beslut om det litterära **Nobelpriset** torde komma att..."

1901-11-16 - sida 2

"...yrkesundervisningen i Tyskland Österrike-Ungern Schweiz Belgien och Holland Det litterära **Nobelpriset** för 1901..."

1901-11-17 - sida 3

"...jordbruk O et litterära **Nobelpriset** Biskop Billing samt professorerna Tegnér Söderwall och Rudin..."

1901-11-30 - sida 2

"...E Åkerberg -i Offentliggörande af vederbörande prl«utdelares beslut angående **Nobelpriset** i fysik..."

1901-12-04 - sida 5

"...af målaremästaren Georg Karlsson Hur C D W förbertdt **Nobelpriset** "Jag föreställer mig att särskildt..."

1901-12-04 - sida 6

"...en litteraturhistoriker som professor Schiick borde akademien med hänsyn till **Nobelpriset** nu mer än..."

1901-12-10 - sida 3

"...hvad Posttidningen för i går har att säga Posttidningen yttrar Det litterära **Nobelpriset** Dtn..."

1901-12-11 - sida 5

"...meddelade beslutet rörande **Nobelpriset** i fysik hvilket skedde i följande tal något lågmäldt hållet..."

1901-12-12 - sida 1

"...Ryssland — Utsikterna för dess genomförande frterpellationen om Wreschenaffären Det litterära **Nobelpriset**..."

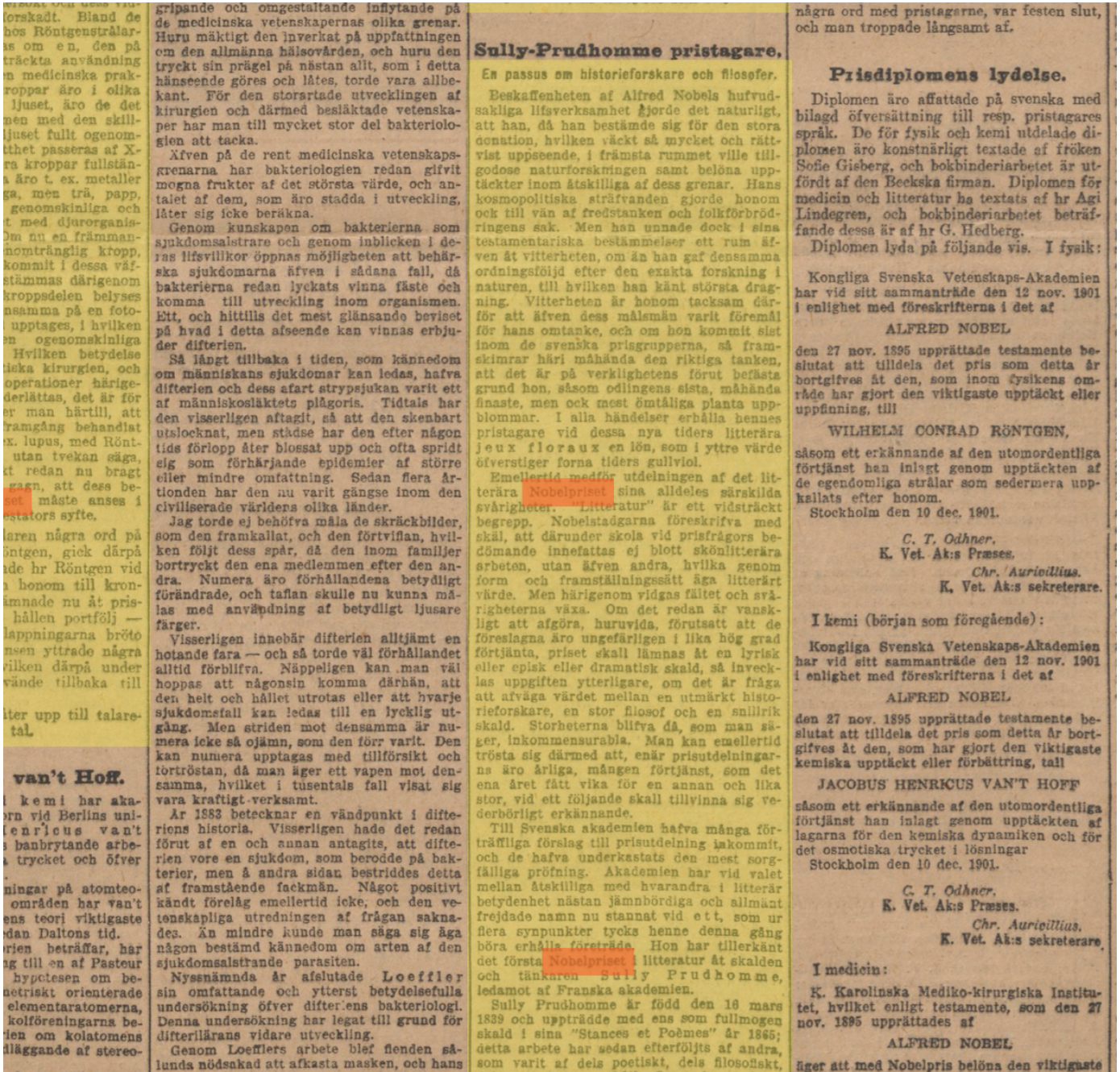
1901-12-12 - sida 3

"...**Nobelpriset** på litteraturens område utan förutfattad mening mot tysk andebildning och tyskt..."

1901-12-14 - sida 1

"...adress till Leo Tolstoj Postverkets stater 1903 Ny svensk Grönlandsexpedition Oet litterära **Nobelpriset**..."

Topic modelling



Sully-Prudhomme pristagare.

En passus om historieforskare och filosofer.

Beskaffenheten af Alfred Nobels hufvudsakliga lifsverksamhet gjorde det naturligt, att han, då han bestämde sig för den stora donation, hvilken väckt så mycket och rättvist uppseende, i främsta rummet ville tillgodose naturforskningen samt belöna upptäckter inom åtskilliga af dess grenar. Hans kosmopolitiska sträfvanden gjorde honom ock till vän af fredstanken och folkförbrödringens sak. Men han unnade dock i sina testamentariska bestämmelser ett rum äfven åt vitterheten, om än han gaf densamma ordningsföljd efter den exakta forskning i naturen, till hvilken han känt största dragning. Vitterheten är honom tacksam därför att äfven dess målsmän varit föremål för hans omtanke, och om hon kommit sist inom de svenska prisgrupperna, så framskimrar häri måhända den riktiga tanken, att det är på verklighetens förut befästa grund hon, såsom odlingens sista, måhända finaste, men ock mest ömtåliga planta uppblommar. I alla händelser erhålla hennes pristagare vid dessa nya tiders litterära jeux floraux en lön, som i yttre värde öfverstiger forna tiders gullviol.

Emellertid medför utdelningen af det litterära Nobelpriset sina alldeles särskilda svårigheter. "Litteratur" är ett vidsträckt begrepp. Nobelstadgarna föreskrifva med skäl, att därunder skola vid prisfrågors bedömande innefattas ej blott skönlitterära arbeten, utan äfven andra, hvilka genom form och framställningssätt äga litterärt värde. Men härigenom vidgas fältet och svårigheterna växa. Om det redan är vanskligt att afgöra, huruvida, förutsatt att de föreslagna äro ungefärligen i lika hög grad förtjänta, priset skall lämnas åt en lyrisk eller episk eller dramatisk skald, så invecklas uppgiften ytterligare, om det är fråga att afväga värdet mellan en utmärkt historieforskare, en stor filosof och en snillrik skald. Storheterna blifva då, som man säger, inkommensurabla. Man kan emellertid trösta sig därmed att, enär prisutdelningarna äro årliga, mången förtjänst, som det ena året fått vika för en annan och lika stor, vid ett följande skall tillvinna sig vederbörligt erkännande.

Till Svenska akademien hafva många förträffliga förslag till prisutdelning inkommit, och de hafva underkastats den mest sorgfälliga pröfning. Akademien har vid valet mellan åtskilliga med hvarandra i litterär betydenhet nästan jämnbördiga och allmänt fröjdade namn nu stannat vid ett, som ur flera synpunkter tycks henne denna gång böra erhålla företräde. Hon har tillerkänt det första Nobelpriset i litteratur åt skalden och tänkaren Sully Prudhomme, ledamot af Franska akademien.

Sully Prudhomme är född den 16 mars 1839 och uppträdde med ens som fullmogen skald i sina "Stances et Poèmes" år 1865; detta arbete har sedan efterföljts af andra, som varit af dels poetiskt, dels filosofiskt,

några ord med pristagarne, var festen slut, och man troppade långsamt af.

Prisdiplomens lydelse.

Diplomen äro affattade på svenska med bilagd öfversättning till resp. pristagares språk. De för fysik och kemi utdelade diplomen äro konstnärligt textade af fröken Sofie Gisberg, och bokbinderiarbetet är utfördt af den Beckska firman. Diplomen för medicin och litteratur ha textats af hr Agi Lindegren, och bokbinderiarbetet beträffande dessa är af hr G. Hedberg.

Diplomen lyda på följande vis. I fysik:

Kongliga Svenska Vetenskaps-Akademien har vid sitt sammanträde den 12 nov. 1901 i enlighet med föreskrifterna i det af

ALFRED NOBEL

den 27 nov. 1895 upprättade testamente beslutat att tilldela det pris som detta år bortgifves åt den, som inom fysikens område har gjort den viktigaste upptäckt eller uppfinning, till

WILHELM CONRAD RÖNTGEN,

såsom ett erkännande af den utomordentliga förtjänst han inlagt genom upptäckten af de egendomliga strålar som sedermera uppkallats efter honom.

Stockholm den 10 dec. 1901.

C. T. Odhner.
K. Vet. Ak:s Præses.

Chr. Aurivillius.
K. Vet. Ak:s sekreterare.

I kemi (början som föregående):

Kongliga Svenska Vetenskaps-Akademien har vid sitt sammanträde den 12 nov. 1901 i enlighet med föreskrifterna i det af

ALFRED NOBEL

den 27 nov. 1895 upprättade testamente beslutat att tilldela det pris som detta år bortgifves åt den, som har gjort den viktigaste kemiska upptäckt eller förbättring, till

JACOBUS HENRICUS VAN'T HOFF

såsom ett erkännande af den utomordentliga förtjänst han inlagt genom upptäckten af lagarna för den kemiska dynamiken och för det osmotiska trycket i lösningar

Stockholm den 10 dec. 1901.

C. T. Odhner.
K. Vet. Ak:s Præses.

Chr. Aurivillius.
K. Vet. Ak:s sekreterare.

I medicin:

K. Karolinska Mediko-kirurgiska Institutet, hvilket enligt testamente, som den 27 nov. 1895 upprättades af

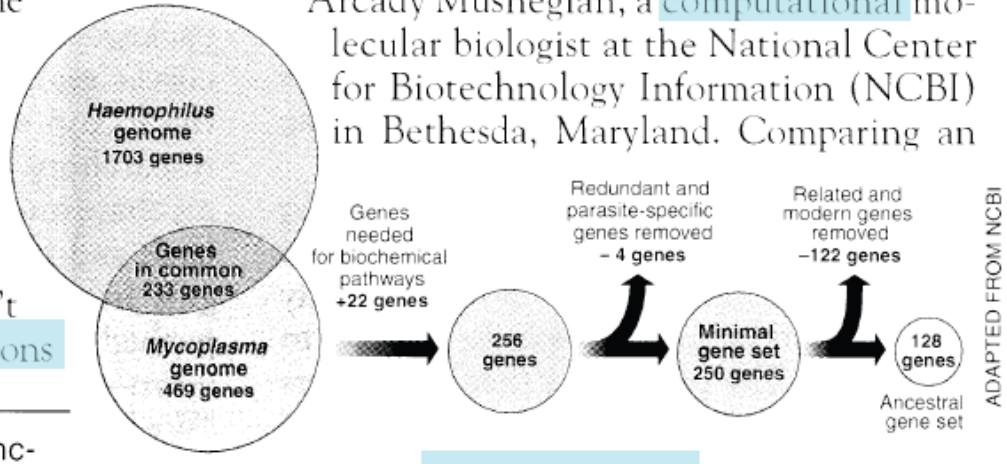
ALFRED NOBEL,

äger att med Nobelpris belöna den viktigaste

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



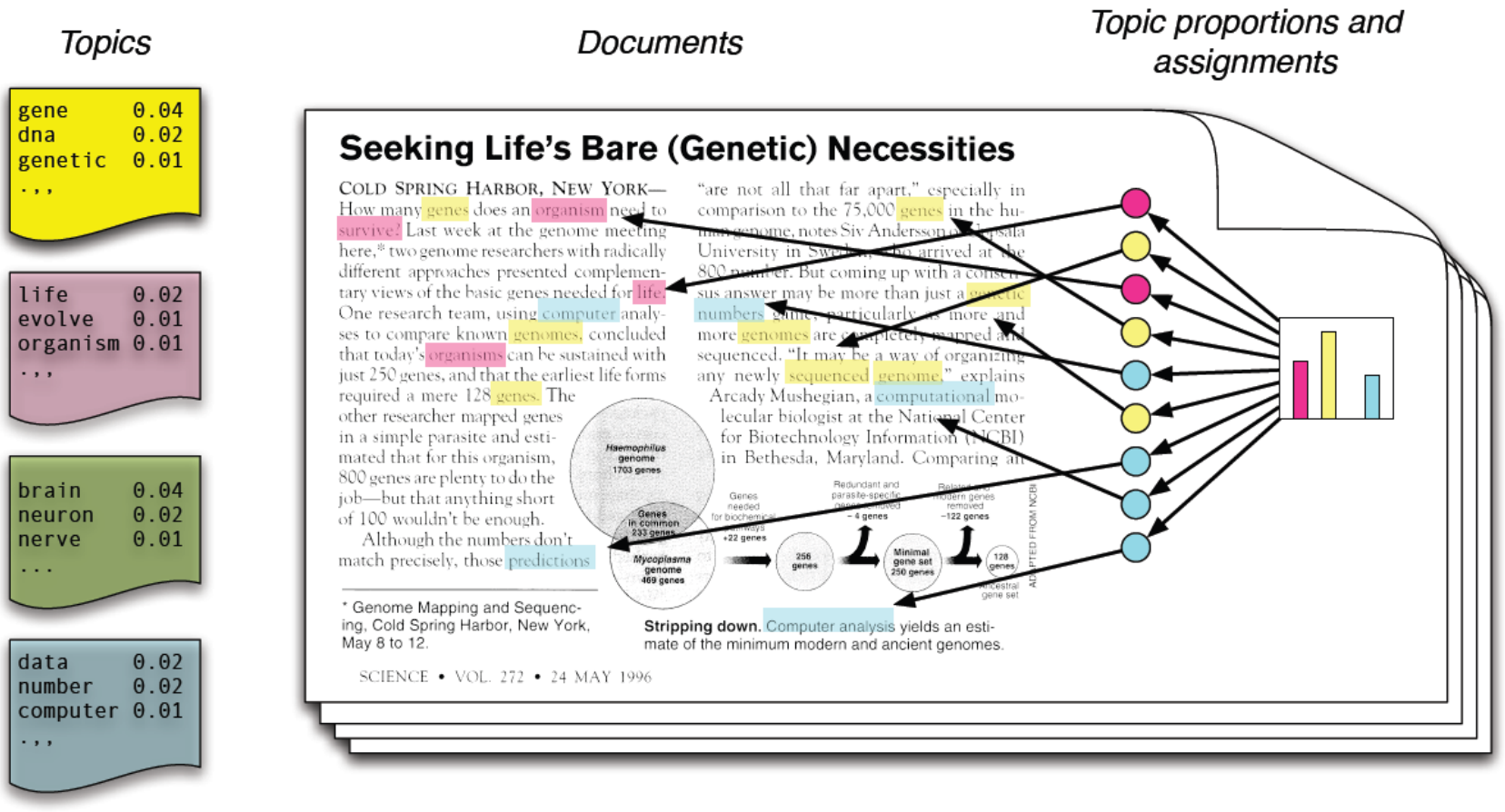
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

1

[1] Credit to D. Blei for the picture

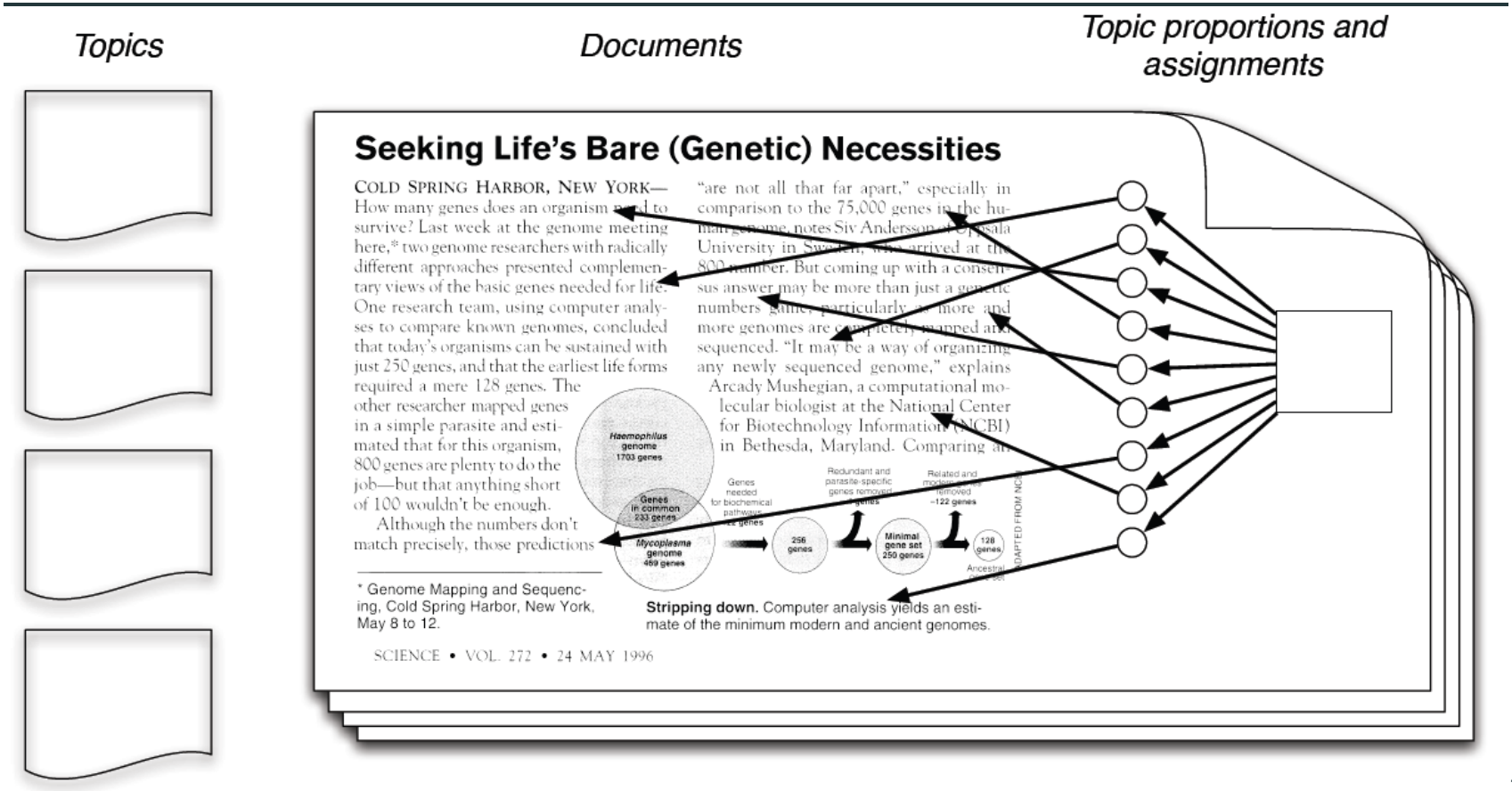
Topic modelling



[1]

[1] Credit to D. Blei for the picture

Topic modelling



1

[1] Credit to D. Blei for the picture

Probabilistic modeling

Directly trying to model topics from data is a hard problem!

Probabilistic modeling

Directly trying to model topics from data is a hard problem!

- In **probabilistic modelling** we constructive a **generative probabilistic process** that could have generated the data.
  In our case we let the topic assignments be a **hidden variable**.

Probabilistic modeling

Directly trying to model topics from data is a hard problem!

- In **probabilistic modelling** we constructive a **generative probabilistic process** that could have generated the data.
  In our case we let the topic assignments be a **hidden variable**.
- We then infer the structure using **posterior inference**.

Probabilistic modeling

Directly trying to model topics from data is a hard problem!

- In **probabilistic modelling** we constructive a **generative probabilistic process** that could have generated the data.
  In our case we let the topic assignments be a **hidden variable**.
- We then infer the structure using **posterior inference**.

Instead of directly modelling the topics, we model how to generate a corpus and then using posterior inference to find the topics.

Probabilistic document modelling

Some notation: We will use bold symbols $\boldsymbol{\theta}$ for the collection of all the random variables, the first sub index θ_i will result in a vector and the second subindex $\theta_{i,j}$ the position in that vector.

We begin by describing the **generative probabilistic process**.

Probabilistic document modelling

Some notation: We will use bold symbols $\boldsymbol{\theta}$ for the collection of all the random variables, the first sub index θ_i will result in a vector and the second subindex $\theta_{i,j}$ the position in that vector.

We begin by describing the **generative probabilistic process**.

1. For each topic, $k = 1, \ldots, K$,

Probabilistic document modelling

Some notation: We will use bold symbols $\boldsymbol{\theta}$ for the collection of all the random variables, the first sub index θ_i will result in a vector and the second subindex $\theta_{i,j}$ the position in that vector.

We begin by describing the **generative probabilistic process**.

1. For each topic, $k = 1, \ldots, K$,
    a. Draw a **distribution over words** for that topic $\beta_k \sim \mathrm{Dir}(\eta)$

Probabilistic document modelling

Some notation: We will use bold symbols $\boldsymbol{\theta}$ for the collection of all the random variables, the first sub index θ_i will result in a vector and the second subindex $\theta_{i,j}$ the position in that vector.

We begin by describing the **generative probabilistic process**.

1. For each topic, $k = 1, \ldots, K$,
   a. Draw a **distribution over words** for that topic $\beta_k \sim \mathrm{Dir}(\eta)$

2. For each document, $d = 1, \ldots, D$

Probabilistic document modelling

Some notation: We will use bold symbols $\boldsymbol{\theta}$ for the collection of all the random variables, the first sub index θ_i will result in a vector and the second subindex $\theta_{i,j}$ the position in that vector.

We begin by describing the **generative probabilistic process**.

1. For each topic, $k = 1, \ldots, K$,
   a. Draw a **distribution over words** for that topic $\beta_k \sim \mathrm{Dir}(\eta)$

2. For each document, $d = 1, \ldots, D$
   a. Draw a vector of **topic proportions** $\theta_d \sim \mathrm{Dir}(\alpha)$

Probabilistic document modelling

Some notation: We will use bold symbols $\boldsymbol{\theta}$ for the collection of all the random variables, the first sub index θ_i will result in a vector and the second subindex $\theta_{i,j}$ the position in that vector.

We begin by describing the **generative probabilistic process**.

1. For each topic, $k = 1, \ldots, K$,
   a. Draw a **distribution over words** for that topic $\beta_k \sim \mathrm{Dir}(\eta)$

2. For each document, $d = 1, \ldots, D$
   a. Draw a vector of **topic proportions** $\theta_d \sim \mathrm{Dir}(\alpha)$

   b. For each **word position**, $n = 1, \ldots, N_d$

Probabilistic document modelling

Some notation: We will use bold symbols $\boldsymbol{\theta}$ for the collection of all the random variables, the first sub index θ_i will result in a vector and the second subindex $\theta_{i,j}$ the position in that vector.

We begin by describing the **generative probabilistic process**.

1. For each topic, $k = 1, \ldots, K$,
   a. Draw a **distribution over words** for that topic $\beta_k \sim \mathrm{Dir}(\eta)$

2. For each document, $d = 1, \ldots, D$
   a. Draw a vector of **topic proportions** $\theta_d \sim \mathrm{Dir}(\alpha)$

   b. For each **word position**, $n = 1, \ldots, N_d$
      i. Draw a **topic assignment** $z_{d,n} \sim \mathrm{Cat}(\theta_d)$

Probabilistic document modelling

Some notation: We will use bold symbols $\boldsymbol{\theta}$ for the collection of all the random variables, the first sub index θ_i will result in a vector and the second subindex $\theta_{i,j}$ the position in that vector.

We begin by describing the **generative probabilistic process**.

1. For each topic, $k = 1, \ldots, K$,
   a. Draw a **distribution over words** for that topic $\beta_k \sim \mathrm{Dir}(\eta)$

2. For each document, $d = 1, \ldots, D$
   a. Draw a vector of **topic proportions** $\theta_d \sim \mathrm{Dir}(\alpha)$

   b. For each **word position**, $n = 1, \ldots, N_d$
      i. Draw a **topic assignment** $z_{d,n} \sim \mathrm{Cat}(\theta_d)$
      ii. Draw a **word** $w_{d,n} \sim \mathrm{Cat}(\beta_{z_{d,n}})$

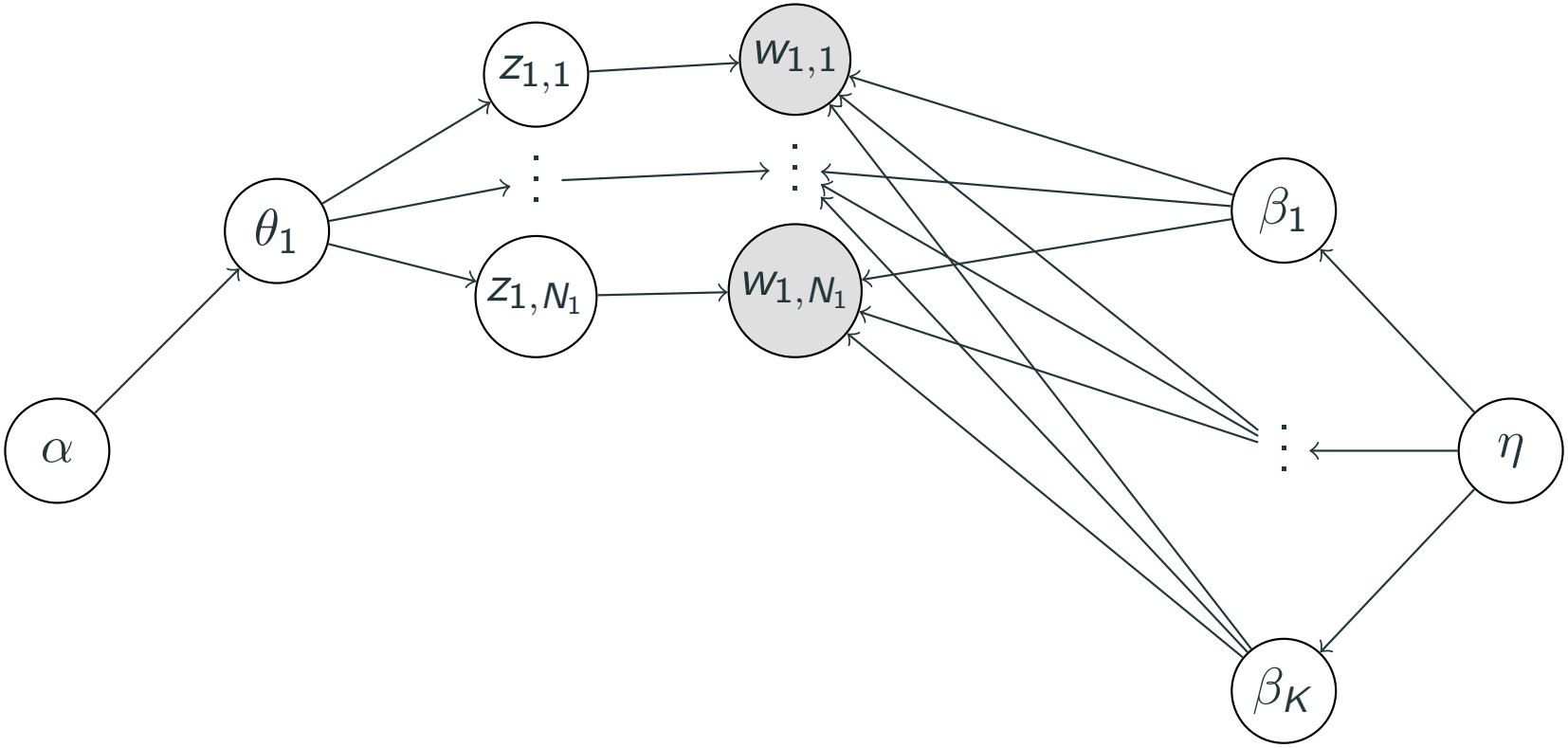
Probabilistic document modelling

Some notation: We will use bold symbols $\boldsymbol{\theta}$ for the collection of all the random variables, the first sub index θ_i will result in a vector and the second subindex $\theta_{i,j}$ the position in that vector.

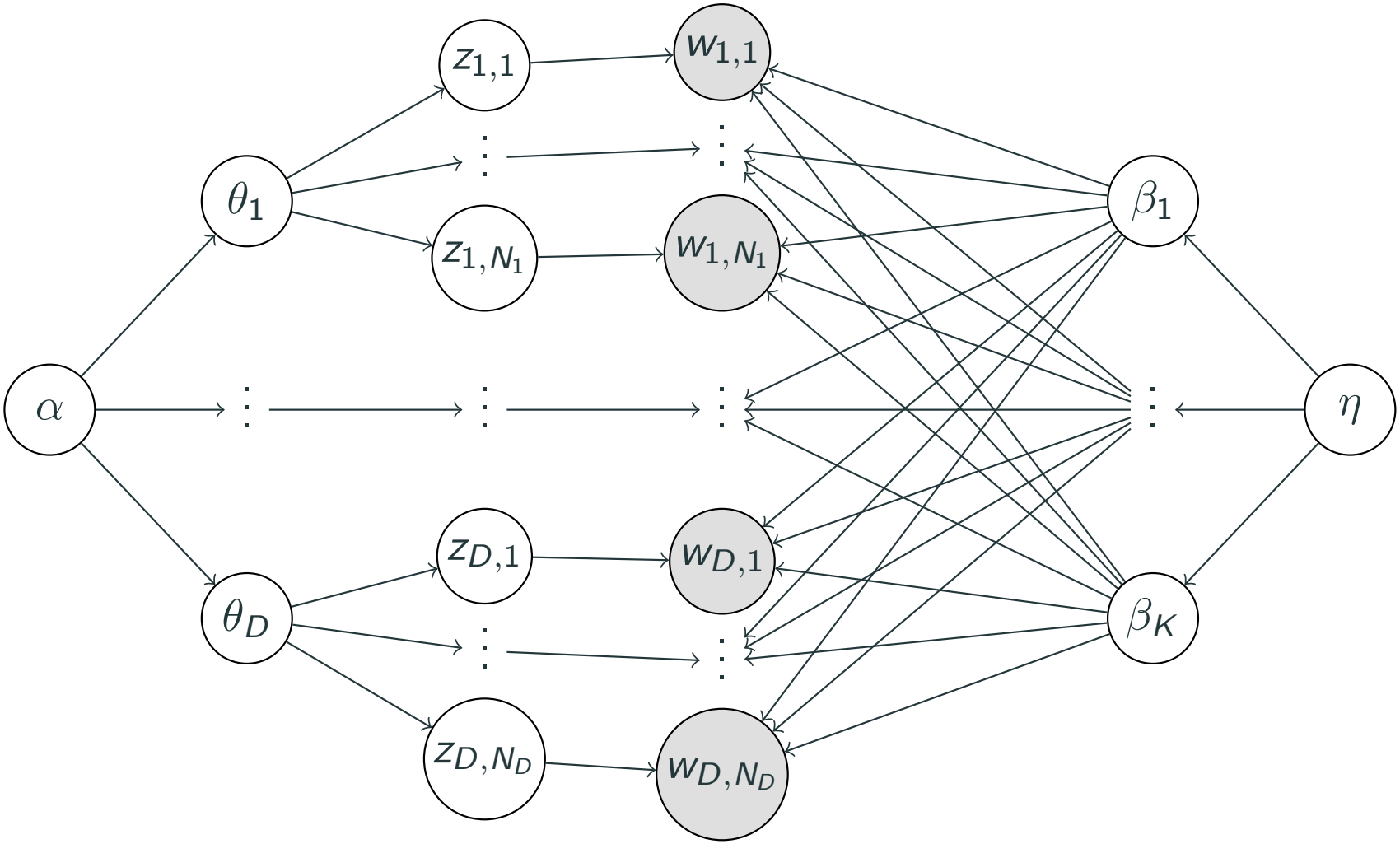
We begin by describing the **generative probabilistic process**.

1. For each topic, $k = 1, \ldots, K$,
   a. Draw a **distribution over words** for that topic $\beta_k \sim \mathrm{Dir}(\eta)$
      — $1 \times V$ **probability vector**
2. For each document, $d = 1, \ldots, D$
   a. Draw a vector of **topic proportions** $\theta_d \sim \mathrm{Dir}(\alpha)$
      — $1 \times K$ **probability vector**
   b. For each **word position**, $n = 1, \ldots, N_d$
      i. Draw a **topic assignment** $z_{d,n} \sim \mathrm{Cat}(\theta_d)$ — **integer** between 1 and K
      ii. Draw a **word** $w_{d,n} \sim \mathrm{Cat}(\beta_{z_{d,n}})$ — **integer** between 1 and V

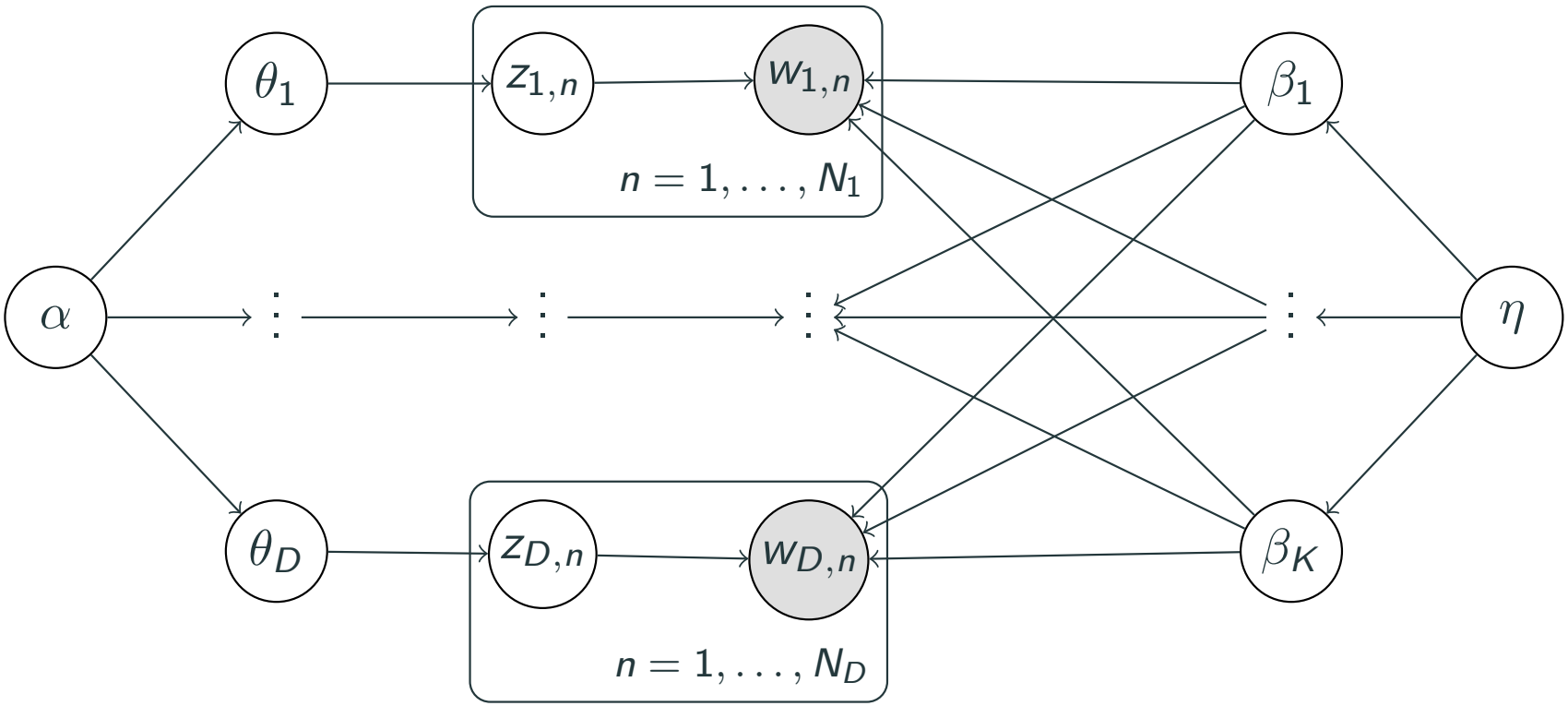
Graphical model



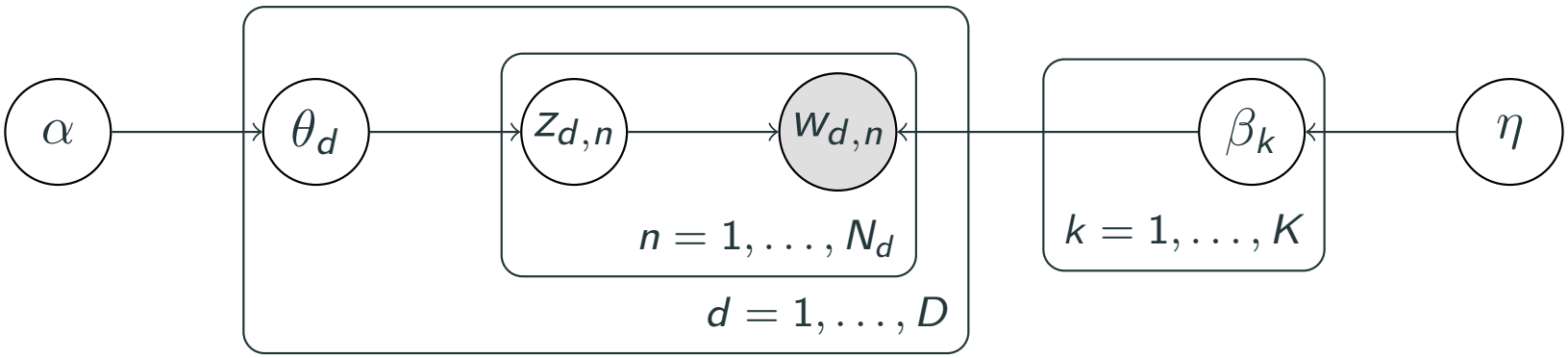
Graphical model



Graphical model



Graphical model



This is known as **plate notation** where each box indicates repetition of those nodes the number of times stated.

The problem

Given a corpus of documents. We wish to infer the distribution of topics on each document and the words within each topic.

The problem

Given a corpus of documents. We wish to infer the distribution of topics on each document and the words within each topic.

Specifically we wish to calculate the **posterior** distribution

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

The problem

Given a corpus of documents. We wish to infer the distribution of topics on each document and the words within each topic.

Specifically we wish to calculate the **posterior** distribution

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

Before going in to how to sample from that distribution, let's see what we can do when we have it. Introduce:

$$\hat{\beta}_k = \mathbb{E}[\beta_k \,|\, \mathbf{w}]$$
$$\hat{\theta}_d = \mathbb{E}[\theta_d \,|\, \mathbf{w}]$$
$$\hat{z}^k_{d,n} = \mathbb{E}[z_{d,n} = k \,|\, \mathbf{w}]$$

With these quantities we can explore the corpus in the following way.

Exploring a corpus

Visualizing a topic: The simplest way to visualize a topic is to look at the estimated probabilities in $\hat{\beta}_k$ and order them.

Exploring a corpus

Visualizing a topic: The simplest way to visualize a topic is to look at the estimated probabilities in $\hat{\beta}_k$ and order them.
A better method, inspired by the **term frequency–inverse document frequency** (TFIDF) score, is the following

$$\text{term-score}_{k,v} = \hat{\beta}_{k,v} \log \left(\frac{\hat{\beta}_{k,v}}{\left(\prod_{j=1}^{K} \hat{\beta}_{j,v} \right)^{1/K}} \right)$$

The first term gives the **term frequency** and is then scaled by weighting down terms that are popular in many topics.

Exploring a corpus

Finding similar documents: Using the **posterior topic proportions** $\hat{\theta}_d$ we have a representation of the document.

To find similar documents we would like to compare these proportions. The Hellinger distance can be used to calculate such a similarity

$$\text{document-similarity}_{d,f} = \sum_{k=1}^{K} \left(\sqrt{\hat{\theta}_{d,k}} - \sqrt{\hat{\theta}_{f,k}} \right)^2 .$$

Exploring a corpus

Finding similar documents: Using the **posterior topic proportions** $\hat{\theta}_d$ we have a representation of the document.
To find similar documents we would like to compare these proportions. The Hellinger distance can be used to calculate such a similarity

$$\text{document-similarity}_{d,f} = \sum_{k=1}^{K} \left(\sqrt{\hat{\theta}_{d,k}} - \sqrt{\hat{\theta}_{f,k}} \right)^2 .$$

Visualizing a document: Finally we would like to use the estimates $\hat{\theta}_d$ and $\hat{z}_{d,n}^k$ to visualize the document.

Exploring a corpus

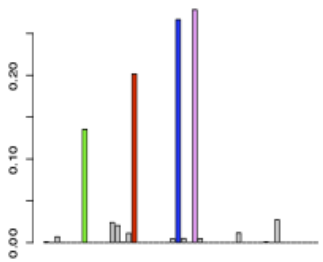
Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions



Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal genera compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

Top Ten Similar Documents

Exhaustive Matching of the Entire Protein Sequence Database
How Big Is the Universe of Exons?
Counting and Discounting the Universe of Exons
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
Ancient Conserved Regions in New Gene Sequences and the Protein Databases
A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure
Testing the Exon Theory of Genes: The Evidence from Protein Structure
Predicting Coiled Coils from Protein Sequences
Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology

2

[2] D. Blei and J. Lafferty. Topic models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis, 2009.

Posterior inference in LDA

We wish to calculate expected values under the posterior distribution

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

Posterior inference in LDA

We wish to calculate expected values under the posterior distribution

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

Looking at the numerator we have

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

Posterior inference in LDA

We wish to calculate expected values under the posterior distribution

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

Looking at the numerator we have

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

$$= \prod_{k=1}^{K} p(\beta_k \,|\, \eta) \prod_{d=1}^{D} \left[p(\theta_d \,|\, \alpha) \prod_{n=1}^{N_d} \left[p(z_{d,n} \,|\, \theta_d) p(w_{d,n} \,|\, z_{d,n}, \beta_{z_{d,n}}) \right] \right]$$

Posterior inference in LDA

We wish to calculate expected values under the posterior distribution

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

Looking at the numerator we have

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

$$= \prod_{k=1}^{K} p(\beta_k \,|\, \eta) \prod_{d=1}^{D} \left[p(\theta_d \,|\, \alpha) \prod_{n=1}^{N_d} \left[p(z_{d,n} \,|\, \theta_d) p(w_{d,n} \,|\, z_{d,n}, \beta_{z_{d,n}}) \right] \right]$$

For the denominator we have

$$p(\mathbf{w} \,|\, \alpha, \eta) = \int \int \sum_{\mathbf{z}} p(\boldsymbol{\beta} \,|\, \eta) p(\boldsymbol{\theta} \,|\, \alpha) p(\mathbf{z} \,|\, \boldsymbol{\theta}) p(\mathbf{w} \,|\, \mathbf{z}, \boldsymbol{\beta}) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\beta},$$

which is intractable.

Gibbs sampling in LDA

We will now look at the different conditionals needed for the Gibbs sampler.

- $\boldsymbol{\theta}$ We notice that the Dirichlet is conjugate prior to the categorical and have that
$$\theta_d \,|\, \boldsymbol{\beta}, \mathbf{z}, \mathbf{w}, \alpha, \eta \sim \mathrm{Dir}(\alpha + c_d)$$
- $\boldsymbol{\beta}$ We notice that the Dirichlet is conjugate prior to the categorical and have that
$$\beta_k \,|\, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, \alpha, \eta \sim \mathrm{Dir}(\eta + \tilde{c}_k)$$
- $\mathbf{z}$ This is a discrete random variable, so we calculate the probabilities directly.
$$p(z_{d,n} = k \,|\, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{w}, \alpha, \eta) \propto \theta_{d,k} \beta_{k,w_{d,n}}$$

Here we have that the elements of c_d and $\tilde{c}_k$ is given by

$$c_{d,k} = \sum_n \mathbb{1}\{z_{d,n} = k\}, \quad \tilde{c}_{k,v} = \sum_d \sum_n \mathbb{1}\{w_{d,n} = v\} \mathbb{1}\{z_{d,n} = k\}$$

Collapsed Gibbs sampling

- Performing Gibbs in this way requires us (at each iteration) to sample from $D \times K + K \times V + \sum_d N_d$ variables. Sampling from such a high-dimensional distribution using Gibbs sampling usually results in poor mixing.

Collapsed Gibbs sampling

- Performing Gibbs in this way requires us (at each iteration) to sample from $D \times K + K \times V + \sum_d N_d$ variables. Sampling from such a high-dimensional distribution using Gibbs sampling usually results in poor mixing.
- To improve the mixing of the algorithm we can **marginalize** some of the variables and sample from the remaining ones.
  - That is instead of sampling from $p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta} \mid \mathbf{w}, \eta, \alpha)$ we only sample from $p(\mathbf{z} \mid \mathbf{w}, \eta, \alpha)$
  - This will help with the mixing.
  - **Introduces dependencies** between the topic variables.

Collapsed Gibbs sampling

- Performing Gibbs in this way requires us (at each iteration) to sample from $D \times K + K \times V + \sum_d N_d$ variables. Sampling from such a high-dimensional distribution using Gibbs sampling usually results in poor mixing.
- To improve the mixing of the algorithm we can **marginalize** some of the variables and sample from the remaining ones.
  - That is instead of sampling from $p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta} \mid \mathbf{w}, \eta, \alpha)$ we only sample from $p(\mathbf{z} \mid \mathbf{w}, \eta, \alpha)$
  - This will help with the mixing.
  - **Introduces dependencies** between the topic variables.
- This can be compared to **Rao-Blackwellisation** where marginalization over variables help with estimation.

Collapsed Gibbs sampling in LDA

In the LDA model we can integrate $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ and just keep the topic choices $\mathbf{z}$. We only need to sample the $\mathbf{z}$ from the resulting marginalized distribution. This is known as **collapsed Gibbs sampling**.

For our collapsed Gibbs sampling we need

$$
\begin{aligned}
&p(z_{d,n} = k \mid \mathbf{z}^{-(d,n)}, \mathbf{w}, \alpha, \eta) \\
&\propto p(z_{d,n} = k \mid \mathbf{z}^{-(d,n)}, \alpha) p(w_{d,n} \mid z_{d,n} = k, \mathbf{w}^{-(d,n)}, \mathbf{z}^{-(d,n)}, \eta) \\
&\propto \frac{\alpha + c_{d,k}^{-(d,n)}}{\sum_{k'=1}^{K} \alpha + c_{d,k'}^{-(d,n)}} \times \frac{\eta + \tilde{c}_{k,w_{d,n}}^{-(d,n)}}{\sum_{v=1}^{V} \eta + \tilde{c}_{k,v}^{-(d,n)}}
\end{aligned}
$$

Given a set of $\mathbf{z}$ we can then estimate $\hat{\theta}_d$ and $\hat{\beta}_k$ by

$$
\hat{\theta}_{d,k} = \frac{\alpha + c_{d,k}}{\sum_{k'=1}^{K} \alpha + c_{d,k'}} \qquad \hat{\beta}_{k,v} = \frac{\eta + \tilde{c}_{k,v}}{\sum_{v'=1}^{V} \eta + \tilde{c}_{k,v'}}
$$

The algorithm

- Initialize $\mathbf{z}$ randomly.
- Create all the counting matrices.
- At every iteration:
  - For every document $d = 1, \ldots, D$ and every word $n = 1, \ldots, N_d$:
    - Set $v = w_{d,n}$
    - Set $k = z_{d,n}$
    - Set $c_{d,k}$ $-= 1$
    - Set $\tilde{c}_{k,v}$ $-= 1$
    - Sample $k \sim \mathrm{Cat}(\{p(z_{d,n} = k' \mid \mathbf{z}^{-(d,n)}, \mathbf{w}, \alpha, \eta)\}_{k'=1}^{K})$
    - Set $c_{d,k}$ $+= 1$
    - Set $\tilde{c}_{k,v}$ $+= 1$

Pre-processing of the documents

Before we can perform this algorithm, some pre-processing of the documents are needed.

Tokenization: Take each document, split it into words. Change all letters to lowercase and remove punctuation.

Small words: Remove words with fewer than 3 characters.

Stopwords: Remove any stopwords. These are some of the most common words for the language.

Lemmatize: Words in third person are changed to first person and verbs are changed into present.

Stem: Words are reduced to their root form.

After this is done you construct your dictionary.

There are may packages that will handle all of these steps for you, no need to reinvent the wheel.

Extensions

- We don't have to work with text documents. There is nothing in the model that stops it from being used in images.

Extensions

- We don't have to work with text documents. There is nothing in the model that stops it from being used in images.
- We can study how topics change over time, by splitting the dataset and constucting one LDA model for each time. The distributions over words β_k are then connected sequentially.

Extensions

- We don't have to work with text documents. There is nothing in the model that stops it from being used in images.
- We can study how topics change over time, by splitting the dataset and constucting one LDA model for each time. The distributions over words β_k are then connected sequentially.
- You can probbly think about something else that can be done with the model.