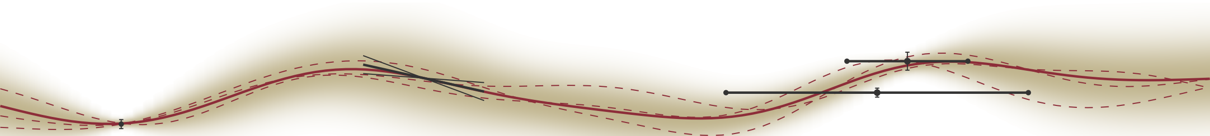
Probabilistic Inference and Learning
Lecture 23
An Extensive Example I

Philipp Hennig

21 January 2019

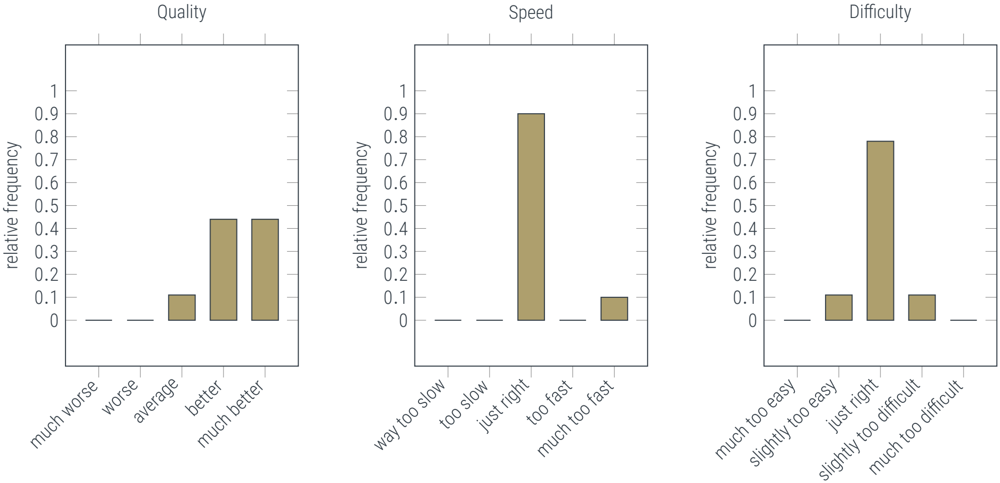
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning



Last Lecture: Debrief

Feedback dashboard



Last Lecture: Debrief

Detailed Feedback

Things you did not like:

- that MH converges slowly :(

Things you did not understand:

- the visualization of MH in the first half
- why not use a Normal-inverse-Wishart prior for the seven scientists?
- why are the tools in the box in this order?
- (how to) apply MC to probabilistic inference

Things you enjoyed:

- the **example**, how to model a problem
- code

0. Introduction to Reasoning under Uncertainty
1. Probabilistic Reasoning
2. Probabilities over Continuous Variables
3. Gaussian Probability Distributions
4. Gaussian Parametric Regression
5. More on Parametric Regression
6. Gaussian Processes
7. More on Kernels & GPs
8. A practical GP example
9. Markov Chains, Time Series, Filtering
10. Classification
11. Empirical Example of Classification
12. Bayesianism and Frequentism
13. Stochastic Differential Equations
14. Exponential Families
15. Graphical Models
16. Factor Graphs
17. The Sum-Product Algorithm
18. Mixture Models
19. The EM Algorithm
20. Variational Inference
21. Monte Carlo
22. Markov Chain Monte Carlo
23. Advanced Modelling Example I
24. Advanced Modelling Example II
25. Advanced Modelling Example III
26. Advanced Modelling Example IV
27. Some Wild Stuff
28. Revision

today’s goal:
Build a Model of History

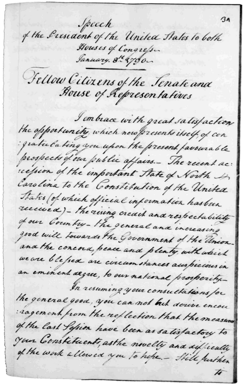
The State of the Union

1790 – 2019

> **[The President] shall from time to time give to the Congress Information of the State of the Union, and recommend to their Consideration such Measures as he shall judge necessary and expedient.**

Article II, §3 of the US Constitution

- Delivered annually since 1790
- Summarizes affairs of the US federal government
- historically delivered in writing, generally spoken since 1982,
- on radio since 1923, TV since 1947, in the evenings since 1965, webcast since 2002
- the inaugural SotU of a new president typically has a different tone



3A

Speech
of the President of the United States to both
Houses of Congreſs
January. 8th 1790

Fellow Citizens of the Senate and
House of Representatives

I embrace with great satisfaction the opportunity, which now presents itself, of congratulating you upon the present favourable prospects of our public affairs. – The recent acceſsion of the important State of North Carolina to the Constitution of the United States (of which official information has been received) – the rising credit and respectability of our Country – the general and increasing good will towards the Government of the Union and the concord, peace and plenty, with which we are bleſsed are circumstances auspicious in an eminent degree, to our national prosperity. –

In resuming your consultations for the general good, you can not but derive encouragement from the reflection that the measures of the last Seſsion have been as satisfactory to your Constituents, as the novelty and difficulty of the work allowed you to hope. – Still further

to

The State of the Union

1790 – 2019

> **[The President] shall from time to time give to the Congress Information of the State of the Union, and recommend to their Consideration such Measures as he shall judge necessary and expedient.**

Article II, §3 of the US Constitution

- Delivered annually since 1790
- Summarizes affairs of the US federal government
- historically delivered in writing, generally spoken since 1982,
- on radio since 1923, TV since 1947, in the evenings since 1965, webcast since 2002
- the inaugural SotU of a new president typically has a different tone

A Flawed but Useful Summary of (US) History

228 years of history

The SotU Addresses are not a perfect reflection of US history, but they are …

- available in their entirety online
- available without interruption for over 200 years
- topical
- given in a reasonably similar setting, annually

Our task: Find **topics** of US history over time.

This is an **unsupervised dimensionality reduction** task.

Nota Bene:

- This is not a course in natural language processing!
- There is an entire toolbox of models for text analysis that will not be discussed here (latent semantic analysis, NNMF, etc.). Some of them have probabilistic interpretation, others don't.
- The point of this exercise is to try out the tools developed in this course on a practical problem. There is no claim that this is the "best" thing to do

However, the model ultimately developed here is likely unusually expressive in its structure, and more flexible than the standard tools. Key takeaway: It does pay to spend time developing your model!

Designing a Probabilistic Machine Learning Model

1. Take a close look at the **Data**
   - do they look reasonable?
   - Any obvious problems? Missing Values, Repetitions?
   - Available Metadata?
2. Think about modelling goals, decide on data types
   - What kind of structure do we seek, what is ok to leave out?

A look at the data

A Look at the Data

explanatory data analysis

note: This is not a NLP course, and certainly not linguistics

- 230 documents (1790 – 2018; 2 in 1961 (Eisenhower & JFK))
- on average $\sim$ 6400 words per speech

A few observations:

- since we are looking to *reduce* complexity, we necessarily have to throw out a bit of structure (assuming human speech isn't massively redundant)
- e.g., usage of word is significant, but its position in the text is not crucial
- neighboring word pairings might be ignored
- there are many redundant **stop words** required for human understanding but carrying only negligible semantic information

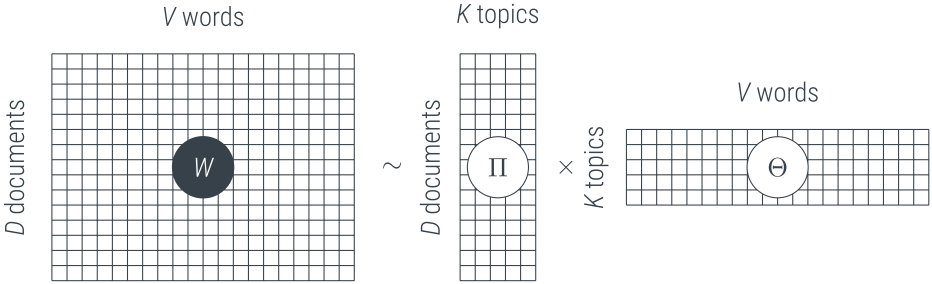
We will model the texts as **Bags of Words**

Designing a Probabilistic Machine Learning Model

1. Take a close look at the **Data**
2. Think about modelling goals, decide on data types
3. Design the **Model**
   - Check variable types
   - Note dependencies in a **directed graph**, study it to note computational bottlenecks
   - use **exponential families** where possible, taking care to use the right family to encode desiderata

Notation and Concept

Towards a Model



We will only store **counts** of words per document, not their order. (Aka. *bags of words*)

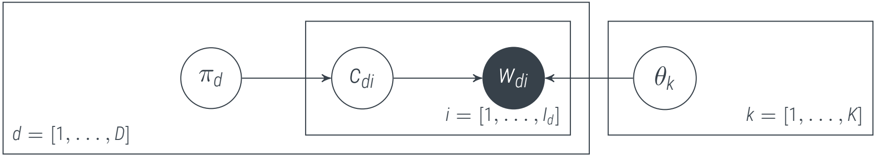
Linear Algebra alone won’t cut it

SVDs are not enough

- The singular value decomposition (SVD) minimizes $\|X - Q\Sigma U'\|_F^2$ for orthonormal matrices $Q \in \mathbb{R}^{D\times K}$ and $U \in \mathbb{R}^{V\times K}$, and a diagonal $\Sigma \in \mathbb{R}^{K\times K}$ with positive diagonal entries (the *singular values*).
- We might naïvely think of Q as a mapping from documents to topics, U' from topics to words, and Σ as the relative strength of topics.
- However, there are several problems:
  - the matrices Q, U returned by the SVD are in general *dense*: Every document contains contributions from *every* topic, and *every* topic involves *all* words.
  - the entries in Q, U, Σ are hard to interpret: They do not correspond to probabilities
  - the entries of Q, U can be *negative*! What does it mean to have a negative topic?

A Preliminary Graph

Just getting a feeling for structure



To draw I_d words $w_{di} \in [1, \dots, V]$ of document $d \in [1, \dots, D]$:

- Asssume K **discrete** topic distributions $\theta_k \in [0,1]^V$ over V words
- For each document d, need distribution $\pi_d \in [0,1]^K$ over K topics
- Draw topic assignments c_{ik} of word w_{id} from $p(C \mid \Pi) = \prod_{i,d,k} \pi_{id}^{c_{dik}}$
- Draw word w_{id} from $p(w_{id} = v \mid c_{di}, \Theta) = \prod_k \theta_{kv}^{c_{dik}}$

Topic Modelling

Desiderata

We need a model with the following properties:

doc sparsity each document *d* should only contain a small number of topics

word sparsity each topic *k* should only contain a small number of the words *v* in the vocabulary

non-negativity a topic can only contribute positively to a document

Reminder: The Dirichlet Distribution

a sparsity prior for probability distribution

$$p(\boldsymbol{x} \mid \boldsymbol{\pi}) = \prod_{i=1}^{n} \pi_{x_i} \qquad x \in \{0; \dots, K\}$$

$$= \prod_{k=1}^{K} \pi_k^{n_k} \qquad n_k := |\{x_i \mid x_i = k\}|$$

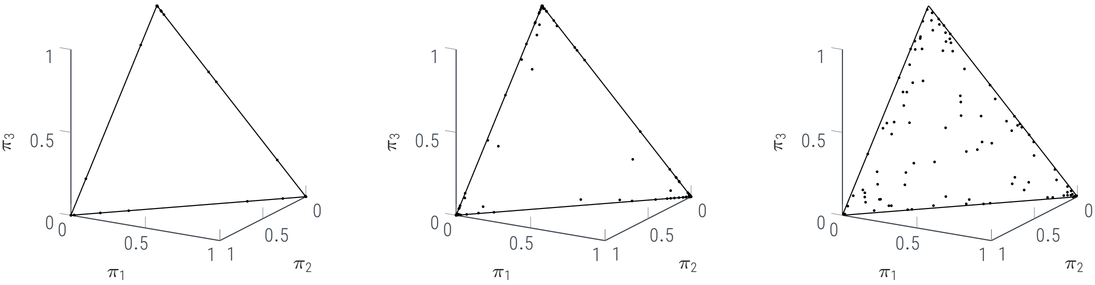
$$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \mathcal{D}(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

$$p(\boldsymbol{\pi} \mid \boldsymbol{x}) = \mathcal{D}(\boldsymbol{\alpha} + \boldsymbol{n})$$

Peter Gustav Lejeune Dirichlet
(1805–1859)

Dirichlets can encode sparsity

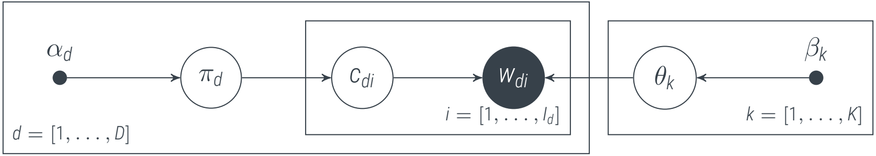
for $\alpha \sim 0.01, 0.1, 0.5$ (100 samples each)



Latent Dirichlet Allocation

Topic Models

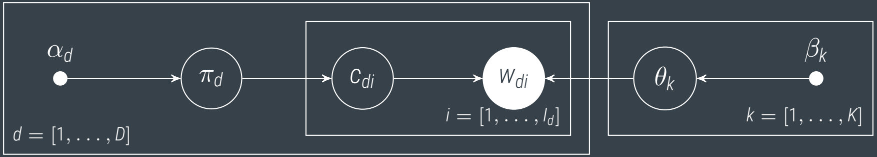
[Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003) JMLR 3, 993–1022]



To draw I_d words $w_{di} \in [1, \dots, V]$ of document $d \in [1, \dots, D]$:

- Draw K topic distributions θ_k over V words from $p(\Theta \mid \boldsymbol{\beta}) = \prod_{k=1}^{K} \mathcal{D}(\theta_k; \beta_k)$
- Draw D document distributions over K topics from $p(\Pi \mid \boldsymbol{\alpha}) = \prod_{d=1}^{D} \mathcal{D}(\pi_d; \alpha_d)$
- Draw topic assignments c_{ik} of word w_{id} from $p(C \mid \Pi) = \prod_{i,d,k} \pi_{id}^{c_{dik}}$
- Draw word w_{id} from $p(w_{id} = v \mid c_{di}, \Theta) = \prod_k \theta_{kv}^{c_{dik}}$

Useful notation: $n_{dkv} = \#\{i : w_{di} = v, c_{ijk} = 1\}$. Write $n_{dk:} := [n_{dk1}, \dots, n_{dkV}]$ and $n_{dk\cdot} = \sum_v n_{dkv}$, etc.



(Directed) Graphical Models provide a visual language to represent ideas and structure. They also allow immediate mathematical insight into **conditional independence**. (Lectures 1, 15)

$$\begin{aligned} p(C, \Pi, \Theta, W) &= \left(\prod_{d=1}^{D} p(\boldsymbol{\pi}_d \mid \boldsymbol{\alpha}_d)\right) \cdot \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} p(c_{di} \mid \pi_d)\right) \cdot \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} p(w_{di} \mid c_{di}, \Theta)\right) \cdot \left(\prod_{k=1}^{K} p(\boldsymbol{\theta}_k \mid \boldsymbol{\beta}_k)\right) \\ &= \left(\prod_{d=1}^{D} \mathcal{D}(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_d)\right) \cdot \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} \left(\prod_{k=1}^{K} \pi_{dk}^{c_{dik}}\right)\right) \cdot \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} \left(\prod_{k=1}^{K} \theta_{kw_{di}}^{c_{dik}}\right)\right) \cdot \left(\prod_{k=1}^{K} \mathcal{D}(\boldsymbol{\theta}_k; \boldsymbol{\beta}_k)\right) \end{aligned}$$

$$p(C,\Pi,\Theta,W) = \left(\prod_{d=1}^{D}\mathcal{D}(\boldsymbol{\pi}_d;\boldsymbol{\alpha}_d)\right)\cdot\left(\prod_{d=1}^{D}\prod_{i=1}^{I_d}\left(\prod_{k=1}^{K}\pi_{dk}^{c_{dik}}\right)\right)\cdot\left(\prod_{d=1}^{D}\prod_{i=1}^{I_d}\left(\prod_{k=1}^{K}\theta_{kw_{di}}^{c_{dik}}\right)\right)\cdot\left(\prod_{k=1}^{K}\mathcal{D}(\boldsymbol{\theta}_k;\boldsymbol{\beta}_k)\right)$$

$$= \left(\prod_{d=1}^{D}\frac{\Gamma(\sum_k \alpha_{dk})}{\prod_k \Gamma(\alpha_{dk})}\prod_{k=1}^{K}\pi_{dk}^{\alpha_{dk}-1+n_{dk\cdot}}\right)\cdot\left(\prod_{k=1}^{K}\frac{\Gamma(\sum_v \beta_{kv})}{\prod_v \Gamma(\beta_{kv})}\prod_{v=1}^{V}\theta_{kv}^{\beta_{kv}-1+n_{\cdot kv}}\right)$$

Exponential Family Distributions are a collection of “standard” distributions for certain data types. Here: for discrete probability distributions (θ,π), the Dirichlet distributions. Exponential families simplify maximum likelihood, MAP and full Bayesian inference by mapping *integration* to *differentiation* and Bayesian inference (through conjugate priors) to *addition* of sufficient statistics. (Lectures 2, 14)

$$p(C, \Pi, \Theta, W) = \left(\prod_{d=1}^{D} \mathcal{D}(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_d)\right) \cdot \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} \left(\prod_{k=1}^{K} \pi_{dk}^{c_{dik}}\right)\right) \cdot \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} \left(\prod_{k=1}^{K} \theta_{kw_{di}}^{c_{dik}}\right)\right) \cdot \left(\prod_{k=1}^{K} \mathcal{D}(\boldsymbol{\theta}_k; \boldsymbol{\beta}_k)\right)$$

$$= \left(\prod_{d=1}^{D} \frac{\Gamma(\sum_k \alpha_{dk})}{\prod_k \Gamma(\alpha_{dk})} \prod_{k=1}^{K} \pi_{dk}^{\alpha_{dk}-1+n_{dk\cdot}}\right) \cdot \left(\prod_{k=1}^{K} \frac{\Gamma(\sum_v \beta_{kv})}{\prod_v \Gamma(\beta_{kv})} \prod_{v=1}^{V} \theta_{kv}^{\beta_{kv}-1+n_{\cdot kv}}\right)$$

Exponential Family Distributions are a collection of “standard” distributions for certain data types. Here: for discrete probability distributions (θ, π), the Dirichlet distributions. Exponential families simplify maximum likelihood, MAP and full Bayesian inference by mapping *integration* to *differentiation* and Bayesian inference (through conjugate priors) to *addition* of sufficient statistics. (Lectures 2, 14)

- Here, the Dirichlet exponential family (the conjugate prior for the multinomial distributions π, θ) **simplifies** the analysis by *collapsing* C out of the joint.
- The Dirichlet is not a “perfectly” expressive language on probabilities (more later). But it allows encoding *sparsity*, which is crucial for this application
- Nevertheless, **as we already know from the graph**, inference on θ, π is tricky, because they are **dependent** (the posteriors both depend on $\boldsymbol{n}$!)