Probabilistic Machine Learning
Lecture 19
Example: Topic Models

Philipp Hennig

28 June 2021

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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

#	date	content	Ex	#	date	content	Ex
1	20.04.	Introduction	1	14	09.06.	Generalized Linear Models	
2	21.04.	Reasoning under Uncertainty		15	15.06.	Exponential Families	8
3	27.04.	Continuous Variables	2	16	16.06.	Graphical Models	
4	28.04.	Monte Carlo		17	22.06.	Factor Graphs	9
5	04.05.	Markov Chain Monte Carlo	3	18	23.06.	The Sum-Product Algorithm	
6	05.05.	Gaussian Distributions		19	29.06.	Example: Modelling Topics	10
7	11.05.	Parametric Regression	4	20	30.06.	Mixture Models	
8	12.05.	Learning Representations		21	06.07.	EM	11
9	18.05.	Gaussian Processes	5	22	07.07.	Variational Inference	
10	19.05.	Understanding Kernels		23	13.07.	Fast Variational Inference	12
11	26.05.	Gauss-Markov Models		24	14.07.	Kernel Topic Models	
12	25.05.	An Example for GP Regression	6	25	20.07.	Outlook	
13	08.06.	GP Classification	7	26	21.07.	Revision	

The Toolbox

Framework:

$$\int p(x_1, x_2)\, dx_2 = p(x_1) \qquad p(x_1, x_2) = p(x_1 \mid x_2)p(x_2) \qquad p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

Modelling:

- graphical models
- Gaussian distributions
- (deep) learnt representations
- Kernels
- Markov Chains
- Exponential Families / Conjugate Priors
- Factor Graphs & Message Passing

Computation:

- Monte Carlo
- Linear algebra / Gaussian inference
- maximum likelihood / MAP
- Laplace approximations
-

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

the goal for (most of) the rest of the course:
Build a Model of History

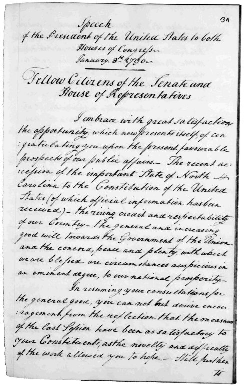
The State of the Union

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



31

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 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.

Disclaimer:

- 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. 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!

Our Goal: Build *craftware*: customized, effective and efficient solution to the learning task.
Use toolboxes where they help, be willing to write our own solution where necessary.

A Look at the Data

explanatory data analysis

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

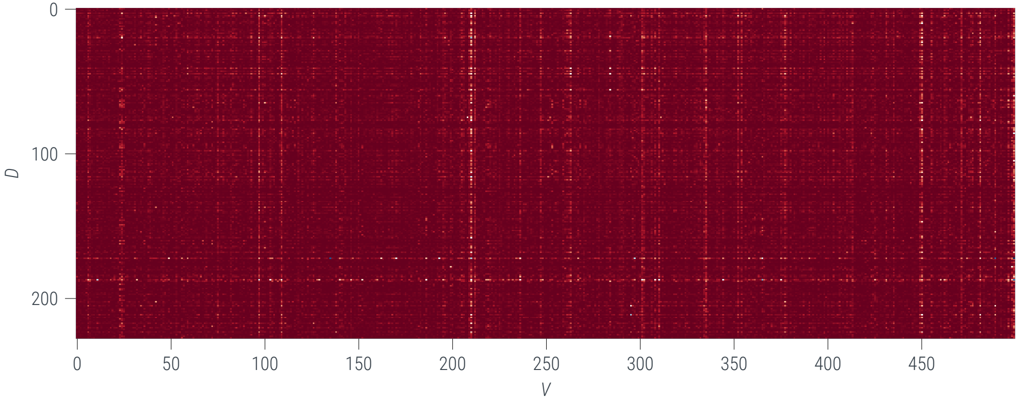
- $D = 231$ documents (1790 – 2019; 2 in 1961 (Eisenhower & JFK))
- individual documents of length $I_d \sim 10^3$ words
- $V \sim 10\,000$ words in vocabulary

A few first simplifications

- there are many redundant **stop words** required for human understanding but carrying only negligible semantic information
- since we are looking to *reduce* complexity, we necessarily have to throw out a bit of structure
- e.g., usage of word is significant, but its position in the text is not crucial. We will model the texts as **Bags of Words**

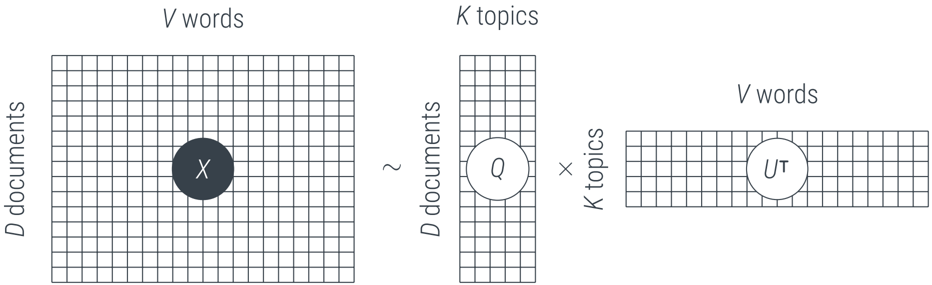
Bags of Words

A look at the data



A Reduced Representation

low-rank decomposition



Consider a dataset $X \in \mathbb{R}^{D \times V}$. **Dimensionality Reduction** aims to find an **encoding** $\phi : \mathbb{R}^V \rightharpoonup \mathbb{R}^K$ and a **decoding** $\psi : \mathbb{R}^K \rightharpoonup \mathbb{R}^V$ with $K \ll V$ such that the encoded representation

$$Z := \phi(X) \in \mathbb{R}^{D \times K}$$

is a *good approximation* of X in the sense that some **reconstruction loss** of $\tilde{X} = \psi(Z)$,

$$\mathcal{L}(X, \psi(Z)) = \mathcal{L}(X, \psi \circ \phi(X))$$

is minimized or small. This may be done, e.g., to

- save memory
- construct a low-dimensional visualization
- "find structure"

Linear dimensionality reduction

The classic derivation of PCA

$$\text{Data: } X \in \mathbb{R}^{D \times V} = [\boldsymbol{x}_1; \dots; \boldsymbol{x}_D].$$

- Consider an orthonormal basis $\{\boldsymbol{u}_i\}_{i=1,\dots,V}$, $\boldsymbol{u}_i^\intercal \boldsymbol{u}_j = \delta_{ij}$. Then

$$\boldsymbol{x}_d = \sum_{i=1}^{V} (\boldsymbol{x}_d^\intercal \boldsymbol{u}_i)\boldsymbol{u}_i =: \sum_{i=1}^{V} \alpha_{di} \boldsymbol{u}_i \qquad X = (XU)U^\intercal$$

- An *approximation* in $K < D$ degrees of freedom is given by any set $(A, \boldsymbol{b}, U)$ as

$$\tilde{\boldsymbol{x}}_d := \sum_{k=1}^{K} a_{dk} \boldsymbol{u}_k + \sum_{\ell=K+1}^{V} b_\ell \boldsymbol{u}_\ell$$

What is the *best* approximation?

The *best* approximation

Empirical Risk Minimization derivation of PCA

Let's find $(A, \boldsymbol{b}, U)$ to minimize the *square empirical risk*

$$J = \frac{1}{D}\sum_{d=1}^{D} \|\boldsymbol{x}_d - \tilde{x}_d\|^2 = \frac{1}{D}\sum_{d=1}^{D}\sum_{v=1}^{V}\left[\boldsymbol{x}_d - \sum_{k=1}^{K} a_{dk}\boldsymbol{u}_k - \sum_{j=K+1}^{V} b_j \boldsymbol{u}_j\right]_v^2$$

First, let's find a_{dk} and b_j: Recall $\sum_j u_{ij}u_{kj} = \delta_{ik}$, use $\bar{\boldsymbol{x}} := \frac{1}{D}\sum_d \boldsymbol{x}_d$, to find

$$\frac{\partial J}{\partial a_{d\ell}} = \frac{2}{D}\sum_{v=1}^{V}\left[\boldsymbol{x}_d - \sum_{k=1}^{K} a_{dk}\boldsymbol{u}_k - \sum_{j=K+1}^{V} b_j \boldsymbol{u}_j\right]_v (-u_{\ell v}) \qquad = \frac{2}{D}(-\boldsymbol{x}_d^\intercal \boldsymbol{u}_\ell) + \frac{2}{D} a_{d\ell} \qquad \stackrel{!}{=} 0$$

$$\frac{\partial J}{\partial b_{\ell}} = \frac{2}{D}\sum_{d=1}^{D}\sum_{v=1}^{V}\left[\boldsymbol{x}_d - \sum_{k=1}^{K} a_{dk}\boldsymbol{u}_k - \sum_{j=K+1}^{V} b_j \boldsymbol{u}_j\right]_v (-u_{\ell v}) \qquad = \frac{2}{D}\sum_{d=1}^{D}(-\boldsymbol{x}_d^\intercal \boldsymbol{u}_\ell) + 2b_\ell \qquad \stackrel{!}{=} 0$$

Thus $a_{dk} = \boldsymbol{x}_d^\intercal \boldsymbol{u}_k$, and $b_j = \bar{\boldsymbol{x}}^\intercal \boldsymbol{u}_j$.

With $a_{dk} = \boldsymbol{x}_d^\intercal \boldsymbol{u}_k$, $b_j = \bar{\boldsymbol{x}}^\intercal \boldsymbol{u}_j$, things simplify:

$$\boldsymbol{x}_d - \tilde{\boldsymbol{x}}_d = \boldsymbol{x}_d - \sum_{k=1}^{K} a_{dk}\boldsymbol{u}_k - \sum_{j=K+1}^{V} b_j \boldsymbol{u}_j = \sum_{\ell=1}^{V} (\boldsymbol{x}_d^\intercal \boldsymbol{u}_\ell)\boldsymbol{u}_\ell - \sum_{k=1}^{K} (\boldsymbol{x}_d^\intercal \boldsymbol{u}_k)\boldsymbol{u}_k - \sum_{j=K+1}^{V} (\bar{\boldsymbol{x}}^\intercal \boldsymbol{u}_j)\boldsymbol{u}_j$$

$$= \sum_{\ell=1}^{K} (\boldsymbol{x}_d^\intercal \boldsymbol{u}_\ell)\boldsymbol{u}_\ell - \sum_{k=1}^{K} (\boldsymbol{x}_d^\intercal \boldsymbol{u}_k)\boldsymbol{u}_k + \sum_{\ell=K+1}^{V} (\boldsymbol{x}_d^\intercal \boldsymbol{u}_\ell)\boldsymbol{u}_\ell - \sum_{j=K+1}^{V} (\bar{\boldsymbol{x}}^\intercal \boldsymbol{u}_j)\boldsymbol{u}_j$$

$$= \sum_{j=K+1}^{V} ((\boldsymbol{x}_d - \bar{\boldsymbol{x}})^\intercal \boldsymbol{u}_j)\boldsymbol{u}_j, \text{ so, with the } \textit{sample covariance matrix } S := \frac{1}{D}\sum_{d=1}^{D} (\boldsymbol{x}_d - \bar{\boldsymbol{x}})(\boldsymbol{x}_d - \bar{\boldsymbol{x}})^\intercal$$

$$J = \frac{1}{D}\sum_{d=1}^{D} \|\boldsymbol{x}_d - \tilde{\boldsymbol{x}}_d\|^2 = \frac{1}{D}\sum_{d=1}^{D}\sum_{j=K+1}^{V} ((\boldsymbol{x}_d - \bar{\boldsymbol{x}})^\intercal \boldsymbol{u}_j)^2 = \frac{1}{D}\sum_{j=K+1}^{V}\sum_{d=1}^{D} \boldsymbol{u}_j^\intercal (\boldsymbol{x}_d - \bar{\boldsymbol{x}})(\boldsymbol{x}_d - \bar{\boldsymbol{x}})^\intercal \boldsymbol{u}_j$$

$$= \sum_{j=K+1}^{V} u_j^\intercal S u_j$$

Maybe we can get away with linear algebra?

Principal Component Analysis — Beltrami, 1873, Jordan, 1874, Pearson, 1901, Schmidt, 1907, Hotelling, 1933, Lanczos, 1950

To find a set of *orthonormal* vectors u_i to minimize the square reconstruction error

$$J = \frac{1}{D}\sum_{d=1}^{D} \|\boldsymbol{x}_d - \tilde{\boldsymbol{x}}_d\|^2 = \sum_{j=K+1}^{V} u_j^\intercal S u_j$$

Choose U as the eigenvectors of the sample covariance $S := \frac{1}{D}\sum_{d=1}^{D}(\boldsymbol{x}_d - \bar{\boldsymbol{x}})(\boldsymbol{x}_d - \bar{\boldsymbol{x}})^\intercal$, and get the *best rank K reconstruction* $\tilde{\boldsymbol{x}}_d$ by setting

$$\tilde{\boldsymbol{x}}_d := \sum_{k=1}^{K} a_{dk}\boldsymbol{u}_k + \sum_{j=K+1}^{V} b_j \boldsymbol{u}_j = \sum_{i=1}^{M}(\boldsymbol{x}_d^\intercal \boldsymbol{u}_i)\boldsymbol{u}_i + \sum_{i=M+1}^{D}(\bar{\boldsymbol{x}}^\intercal \boldsymbol{u}_i)\boldsymbol{u}_i$$

This yields $J = \sum_{j=K+1}^{V} \lambda_j$ (where λ_j are the eigenvalues of S, sorted descendingly). If we first center the data $\hat{X} = X - \mathbf{1}\bar{\boldsymbol{x}}^\intercal$, so $\boldsymbol{b} = 0$, the U are the (right) **singular vectors** of $\hat{X} = Q\Sigma U^\intercal$.

Probabilistic PCA

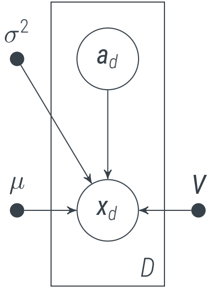
a maximum-likelihood derivation

[Tipping & Bishop, 1997,1999. Sam Roweis, 1998]

Treat the loss, up to scaling, as a non-normalised negative log likelihood:

$$J = -c \cdot \log p(X \mid \tilde{X}) + \log Z = \frac{1}{D} \sum_{d=1}^{D} \|\boldsymbol{x}_d - \tilde{\boldsymbol{x}}_d\|^2$$

$$\Rightarrow p(X \mid \tilde{X}) = \prod_{d=1}^{D} \mathcal{N}(\boldsymbol{x}_d; \tilde{\boldsymbol{x}}_d, \sigma^2 I)$$



We also need to encode that we want a *low-dimensional, linear* embedding, and that the embedding should be in terms of *independent* (orthogonal) dimensions.

Thus, consider

$$\boldsymbol{x}_d = \boldsymbol{V}\boldsymbol{a}_d + \boldsymbol{\mu} + \varepsilon \quad \text{with } p(\boldsymbol{a}_d) = \mathcal{N}(0; I_K), \boldsymbol{V} \in \mathbb{R}^{V \times K} \text{ and } p(\varepsilon) = \mathcal{N}(0; \sigma^2)$$

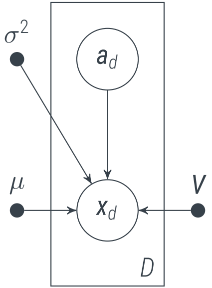
with marginal likelihood (where $C := \boldsymbol{V}\boldsymbol{V}^\intercal + \sigma^2 I$)

$$p(X) = \int \prod_{d=1}^{D} p(\boldsymbol{x}_d \mid \boldsymbol{a}_d) p(\boldsymbol{a}_d)\, d\boldsymbol{a}_d = \prod_d \mathcal{N}(\boldsymbol{x}_d; \boldsymbol{\mu}, C)$$

$$\log p(X) = -\frac{DV}{2}\log(2\pi) - \frac{D}{2}\log|C| - \frac{1}{2}\sum_{d=1}^{D}(\boldsymbol{x}_d - \boldsymbol{\mu})^\intercal C^{-1}(\boldsymbol{x}_d - \boldsymbol{\mu})$$

$$\bar{\boldsymbol{x}} = \arg\max_{\boldsymbol{\mu}} \log p(X), \quad \text{thus the max. lik. can be written as}$$

$$\log p(X) = -\frac{D}{2}\left(V\log(2\pi) + \log|C| + \operatorname{tr}(C^{-1}S)\right)$$



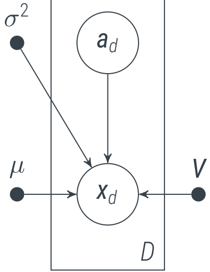
$$\log p(X) = -\frac{D}{2}\left(V\log(2\pi) - \log|C| + \mathrm{tr}(C^{-1}S)\right)$$

yields max. lik. for $\boldsymbol{V}, \sigma^2$ at [Tipping & Bishop, 1999], with $RR^\intercal = I_K$ and $S = U\Lambda U^\intercal$

$$\boldsymbol{V}_{ML} = U_{1:K}(\Lambda_K - \sigma^2 I)^{1/2}R \quad \text{and} \quad \sigma^2_{ML} = \frac{1}{V-K}\sum_{j=K+1}^{V}\lambda_j$$

setting $\sigma^2, \boldsymbol{\mu}, U$ this way, and $R = I$ w.l.o.g., gives posterior

$$\begin{aligned} p(\boldsymbol{a}_d \mid \boldsymbol{x}_d) &= \mathcal{N}(\boldsymbol{a}_d; (\boldsymbol{V}^\intercal \boldsymbol{V} + \sigma^2 I)^{-1}\boldsymbol{V}^\intercal(\boldsymbol{x}_d - \bar{\boldsymbol{x}}), \sigma^2(\boldsymbol{V}^\intercal\boldsymbol{V} + \sigma^2 I)^{-1}) \\ &= \mathcal{N}(\boldsymbol{a}_d; \Lambda_K^{-1}(\Lambda_K - \sigma^2 I_K)^{1/2}U_{1:K}(\boldsymbol{x}_d - \bar{\boldsymbol{x}}), \sigma^2\Lambda^{-1}) \end{aligned}$$



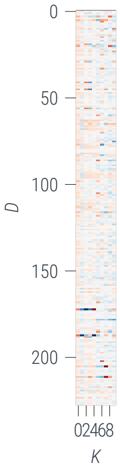
So, does it work?

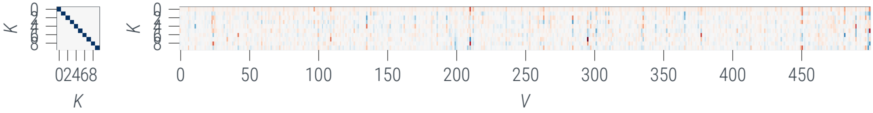
```
count_vect_lsa = CountVectorizer(max_features=VOCAB_SIZE, stop_words=['000'])
X_count = count_vect_lsa.fit_transform(preprocessed).toarray()

U_, S_, V_T_ = np.linalg.svd(X_count, full_matrices=False)
```

Latent Semantic Indexing / Principal Component Analysis

a first result on our dataset





1. tonight fight taxis faith century today enemy fellow
2. year program world new work need help america
3. dollar war program fiscal year expenditure million united
4. man law dollar business national corporation legislation labor
5. administration policy energy program continue development provide effort
6. war nation power man mexico world peace public
7. united war states american world mexico man nation
8. government people states world free shall dollar constitution
9. year free nation world increase report subject great
10. world free gold government bank note american treasury

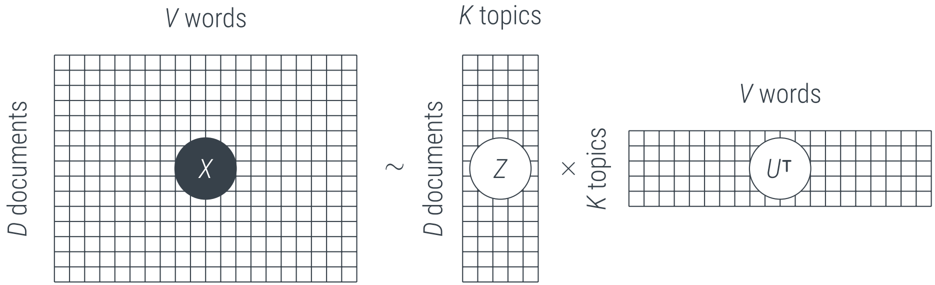
Linear Algebra alone won't cut it

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?

We need Sparsity

How about *one* topic per document?

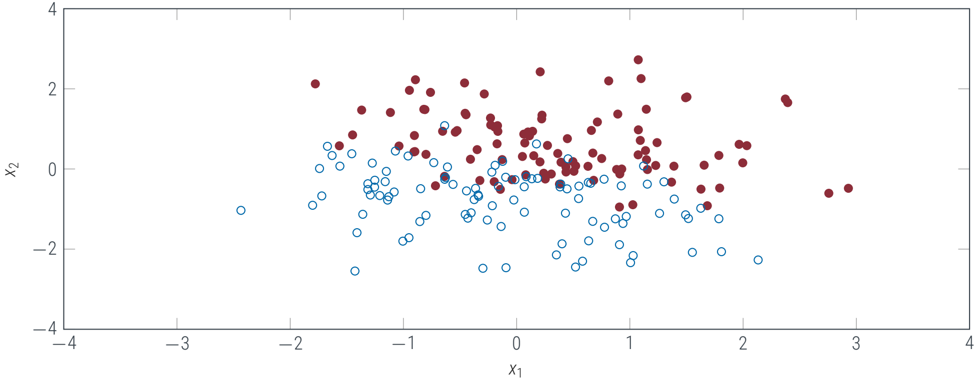


For PCA, we allowed $Z \in \mathbb{R}^{D\times K}$. Maybe we need $Z \in \{0;1\}^{D\times K}$ and $Z\mathbf{1}_K = \mathbf{1}_D$?

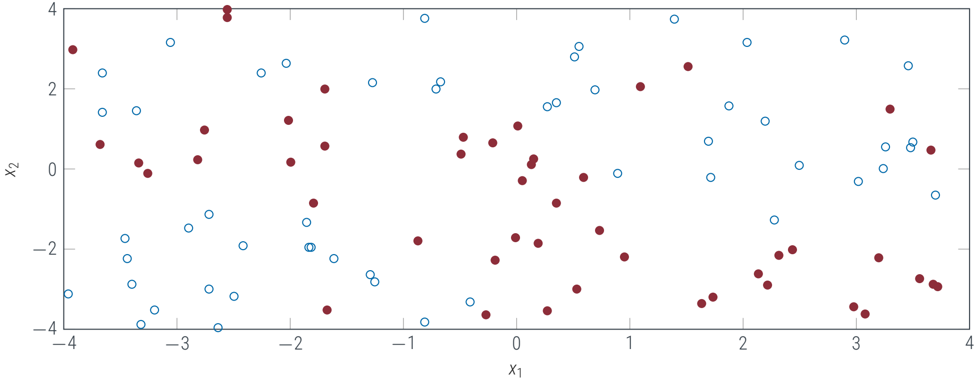
Mixture Models

generative modelling with discrete classes

a **supervised** problem that can be solved **discriminatively** in a *linear* fashion



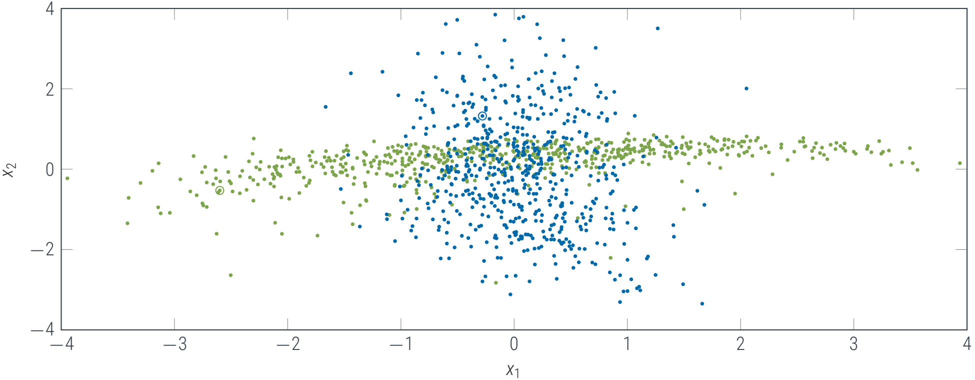
a **supervised** problem that can be solved **discriminatively** in a *nonlinear* fashion



Mixture Models

generative modelling with discrete classes

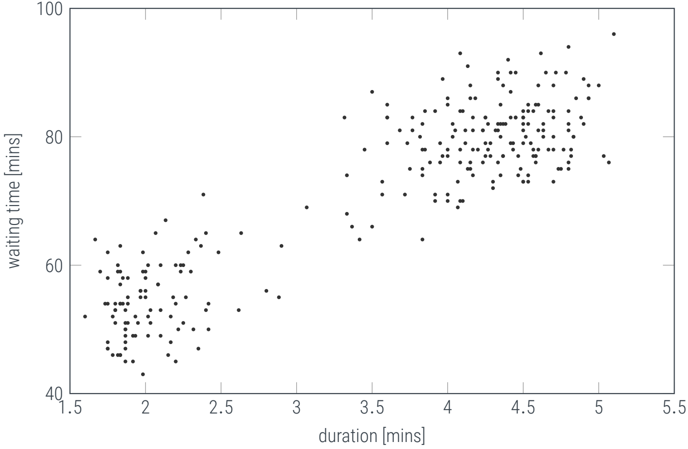
a **supervised** problem that can be solved **generatively** (in a Gaussian fashion?)



an **unsupervised** problem

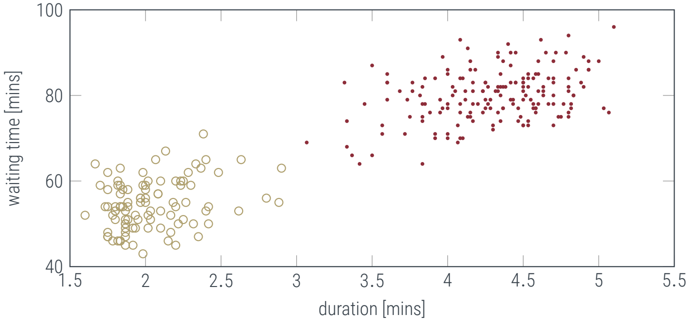
`https://www.stat.cmu.edu/ larry/all-of-statistics/=data/faithful.dat`

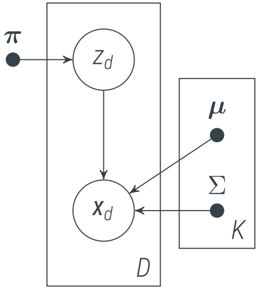
Azzalini, A. and Bowman, A. W. (1990). *A look at some data on the Old Faithful geyser.* Applied Statistics 39, 357-365.



a Gaussian mixture

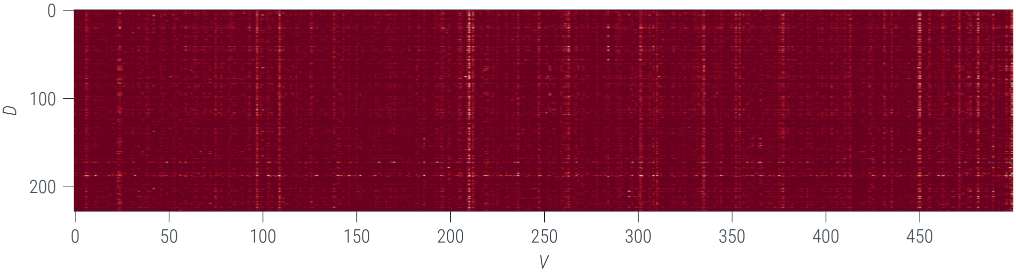
$$p(\boldsymbol{x}_d, Z) = \prod_{d=1}^{D} p(z_d \mid \pi) p(\boldsymbol{x}_d \mid z_d, \boldsymbol{\mu}, \Sigma) = \prod_{d=1}^{D} \prod_{k=1}^{K} \pi_k^{z_{dk}} \mathcal{N}(w_d; \boldsymbol{\mu}_k, \Sigma_k)^{z_{dk}}$$

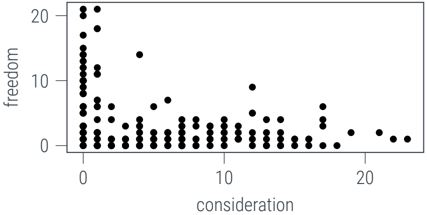


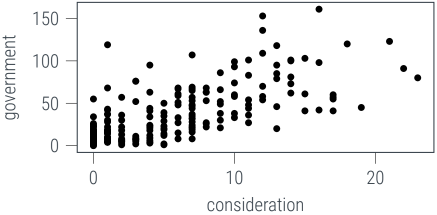


A Gaussian Mixture isn't quite right yet

word counts aren't real-valued

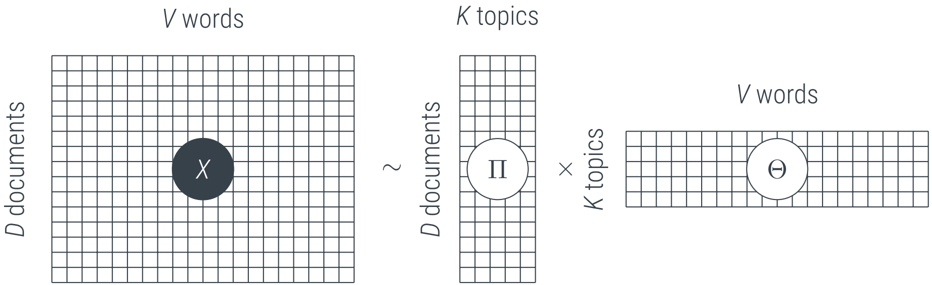






A Mixture of Probabilities?

Desiderata



- topics should be probabilities: $p(\boldsymbol{x}_d \mid k) = \prod_{v=1}^{V} \theta_{kv}^{x_{dv}}$
- but documents should have *several* topics! Let π_{dk} be the *probability* to draw a word from topic k