Clustering

Cédric Archambeau
cedrica@amazon.com

Peyresq Summer School
France, July 2016

Overview

1. Classification (2.5 hours)
2. Clustering (1.5 hours)
3. Practical sessions (1 hour)

Overview

1. Classification (2.5 hours)
2. Clustering (1.5 hours)
3. Practical sessions (1 hour)

LEARNING GOALS

- Understand the difference between clustering and classification
- Understand when to apply clustering
- Understand the EM algorithm
- Being able to derive the EM updates of a mixture models

Overview

1. Classification (2.5 hours)
2. Clustering (1.5 hours)
3. Practical sessions (1 hour)

LEARNING GOALS

- Understand the difference between clustering and classification
- Understand when to apply clustering
- Understand the EM algorithm
- Being able to derive the EM updates of a mixture models
- **Being able to learn by yourself!**

Outline

Outline

What is clustering

- The goal is to identify some structure in the .data
- Typically groups of data points sharing same properties

What is clustering

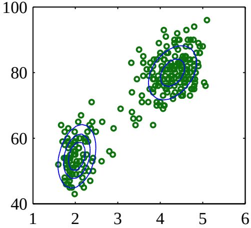
- The goal is to identify some structure in the .data
- Typically groups of data points sharing same properties
- Falls into unsupervised learning bucket (as opposed to classification)

What is clustering

- The goal is to identify some structure in the .data
- Typically groups of data points sharing same properties
- Falls into unsupervised learning bucket (as opposed to classification)
- Discovered structure is based on some strong assumptions about the data

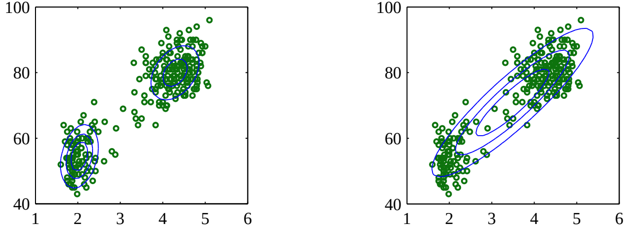
What is clustering

- The goal is to identify some structure in the .data
- Typically groups of data points sharing same properties
- Falls into unsupervised learning bucket (as opposed to classification)
- Discovered structure is based on some strong assumptions about the data



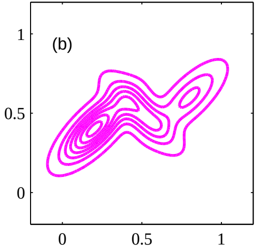
What is clustering

- The goal is to identify some structure in the .data
- Typically groups of data points sharing same properties
- Falls into unsupervised learning bucket (as opposed to classification)
- Discovered structure is based on some strong assumptions about the data

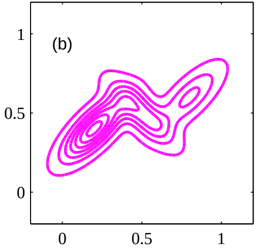


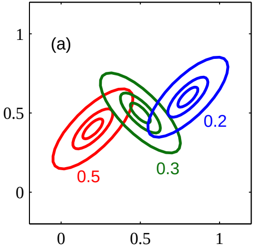
Outline

Mixture models

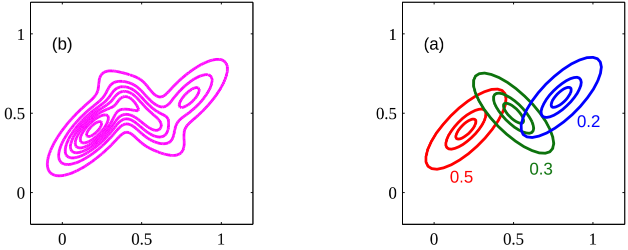


Mixture models





Mixture models



$$p(\boldsymbol{x}) = \sum_k \pi_k p_{\boldsymbol{\theta}_k}(\boldsymbol{x}), \qquad \sum_k \pi_k = 1, \qquad \pi \geqslant 0.$$

Mixture of Gaussians

$$p_{\boldsymbol{\theta}_k}(\boldsymbol{x}) = \mathrm{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- How shall we learn the parameters?

Mixture of Gaussians

$$p_{\boldsymbol{\theta}_k}(\boldsymbol{x}) = \mathrm{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- How shall we learn the parameters?
- By maximimum likelihood?

$$\ln \prod_i p(\boldsymbol{x}_i) = \sum_i \ln \sum_k \pi_k \mathrm{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Mixture of Gaussians

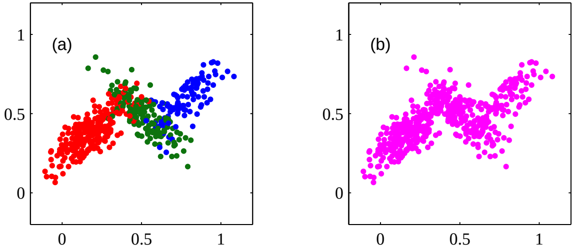
$$p_{\boldsymbol{\theta}_k}(\boldsymbol{x}) = \mathrm{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- How shall we learn the parameters?
- By maximimum likelihood?

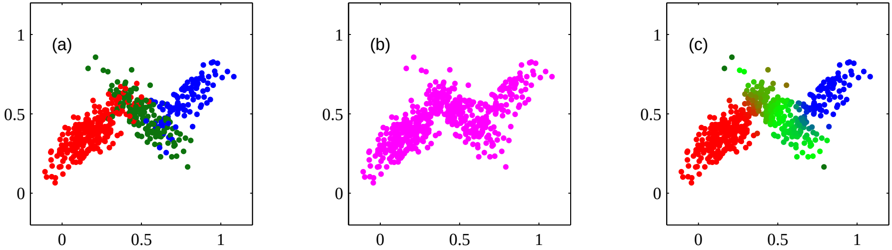
$$\ln \prod_i p(\boldsymbol{x}_i) = \sum_i \ln \sum_k \pi_k \mathrm{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- No closed form solution :-(

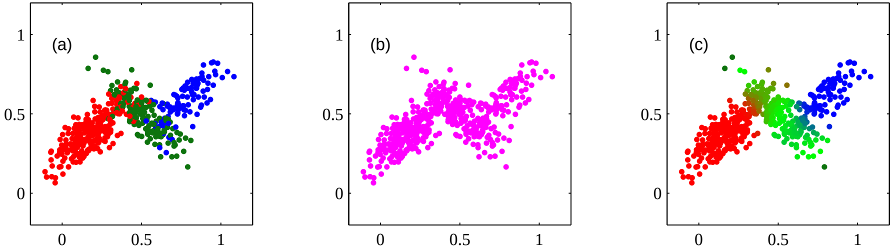
Mixture models: latent variable view



Mixture models: latent variable view

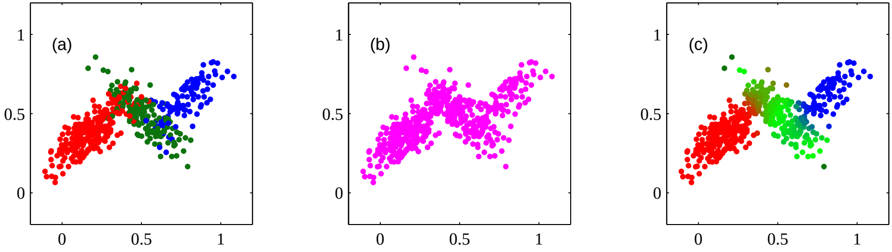


Mixture models: latent variable view



$$p(\boldsymbol{x}|z) = \text{Gaussian}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z), \qquad P(z) = \text{Categorical}\,(\boldsymbol{\pi})\,.$$

Mixture models: latent variable view



$$p(\boldsymbol{x}|z) = \text{Gaussian}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z), \qquad P(z) = \text{Categorical}(\boldsymbol{\pi}).$$

Do we recover the original model?

$$p(\boldsymbol{x}) = \sum_z P(z)p(\boldsymbol{x}|z) = \sum_k \pi_k \text{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Some definitions

- The differential entropy is defined as

$$\mathrm{H}[p(\boldsymbol{x})] = -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) \; d\boldsymbol{x}.$$

The entropy of a Gaussian random variable is given by $\frac{D}{2} \ln 2\pi e + \frac{1}{2} \ln |\boldsymbol{\Sigma}|$.

Some definitions

- The differential entropy is defined as

$$\mathrm{H}[p(\boldsymbol{x})] = -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x})\, d\boldsymbol{x}.$$

  The entropy of a Gaussian random variable is given by $\frac{D}{2} \ln 2\pi e + \frac{1}{2} \ln |\boldsymbol{\Sigma}|$.
- The Kullback-Leibler divergence measures the difference between two densities:

$$\mathrm{KL}[q\|p] = \int q(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\, d\boldsymbol{x} \geqslant 0.$$

  The KL is asymmetric (thus not a distance) and only zero if $q(\boldsymbol{x}) = p(\boldsymbol{x})$ for all $\boldsymbol{x}$.

Expectation-Maximisation (EM)

The EM algorithm maximises a lower bound to the log-marginal likelihood (in presence of latent variables, like parameters):

Expectation-Maximisation (EM)

The EM algorithm maximises a lower bound to the log-marginal likelihood (in presence of latent variables, like parameters):

- Using *Jensen's inequality*, we get for a distribution $q(\boldsymbol{Z})$ within a tractable family:

$$\begin{aligned}
\ln p(\boldsymbol{x}|\boldsymbol{\theta}) &= \ln \int p(\boldsymbol{x}, \boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z} \\
&\geqslant \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{x}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \\
&\equiv -\mathcal{F}(q, \boldsymbol{\theta}).
\end{aligned}$$

Expectation-Maximisation (EM)

The EM algorithm maximises a lower bound to the log-marginal likelihood (in presence of latent variables, like parameters):

- Using *Jensen's inequality*, we get for a distribution $q(\boldsymbol{Z})$ within a tractable family:

$$\begin{aligned} \ln p(\boldsymbol{x}|\boldsymbol{\theta}) &= \ln \int p(\boldsymbol{x}, \boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z} \\ &\geqslant \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{x}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \\ &\equiv -\mathcal{F}(q, \boldsymbol{\theta}). \end{aligned}$$

- The quantity $\mathcal{F}(q, \boldsymbol{\theta})$ can be interpretted as the (variational) free energy from statistical physics.

EM algorithm

The variational free energy $\mathcal{F}(q, \boldsymbol{\theta})$ can be decomposed into two different ways:

$$-\mathcal{F}(q, \boldsymbol{\theta}) = \ln p(\boldsymbol{x}|\boldsymbol{\theta}) - \mathrm{KL}[q(\boldsymbol{Z}) \| p(\boldsymbol{Z}|\boldsymbol{x}, \boldsymbol{\theta})], \qquad \textbf{(E step)}$$

$$-\mathcal{F}(q, \boldsymbol{\theta}) = \langle \ln p(\boldsymbol{x}, \boldsymbol{Z}|\boldsymbol{\theta}) \rangle_{q(\boldsymbol{Z})} + \mathrm{H}[q(\boldsymbol{Z})]. \qquad \textbf{(M step)}$$

EM algorithm

The variational free energy $\mathcal{F}(q, \boldsymbol{\theta})$ can be decomposed into two different ways:

$$-\mathcal{F}(q, \boldsymbol{\theta}) = \ln p(\boldsymbol{x}|\boldsymbol{\theta}) - \mathrm{KL}[q(\boldsymbol{Z}) \| p(\boldsymbol{Z}|\boldsymbol{x}, \boldsymbol{\theta})], \qquad \textbf{(E step)}$$

$$-\mathcal{F}(q, \boldsymbol{\theta}) = \langle \ln p(\boldsymbol{x}, \boldsymbol{Z}|\boldsymbol{\theta}) \rangle_{q(\boldsymbol{Z})} + \mathrm{H}[q(\boldsymbol{Z})]. \qquad \textbf{(M step)}$$

- EM maximises the lower bound by alternating between these two steps; it converges to local optimum of $\ln p(\boldsymbol{x}|\boldsymbol{\theta})$.

EM algorithm

The variational free energy $\mathcal{F}(q, \boldsymbol{\theta})$ can be decomposed into two different ways:

$$-\mathcal{F}(q, \boldsymbol{\theta}) = \ln p(\boldsymbol{x}|\boldsymbol{\theta}) - \mathrm{KL}[q(\boldsymbol{Z}) \| p(\boldsymbol{Z}|\boldsymbol{x}, \boldsymbol{\theta})], \qquad \textbf{(E step)}$$

$$-\mathcal{F}(q, \boldsymbol{\theta}) = \langle \ln p(\boldsymbol{x}, \boldsymbol{Z}|\boldsymbol{\theta}) \rangle_{q(\boldsymbol{Z})} + \mathrm{H}[q(\boldsymbol{Z})]. \qquad \textbf{(M step)}$$

- EM maximises the lower bound by alternating between these two steps; it converges to local optimum of $\ln p(\boldsymbol{x}|\boldsymbol{\theta})$.
- By construction, the EM algorithm ensures a **monotonic** increase of the bound.

EM algorithm

The variational free energy $\mathcal{F}(q, \boldsymbol{\theta})$ can be decomposed into two different ways:

$$-\mathcal{F}(q, \boldsymbol{\theta}) = \ln p(\boldsymbol{x}|\boldsymbol{\theta}) - \mathrm{KL}[q(\boldsymbol{Z}) \| p(\boldsymbol{Z}|\boldsymbol{x}, \boldsymbol{\theta})], \qquad \textbf{(E step)}$$

$$-\mathcal{F}(q, \boldsymbol{\theta}) = \langle \ln p(\boldsymbol{x}, \boldsymbol{Z}|\boldsymbol{\theta}) \rangle_{q(\boldsymbol{Z})} + \mathrm{H}[q(\boldsymbol{Z})]. \qquad \textbf{(M step)}$$

- EM maximises the lower bound by alternating between these two steps; it converges to local optimum of $\ln p(\boldsymbol{x}|\boldsymbol{\theta})$.
- By construction, the EM algorithm ensures a **monotonic** increase of the bound.
- Still ok if q is a good approximation of the true posterior (approximate E step).

EM algorithm

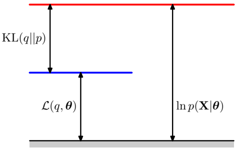
The variational free energy $\mathcal{F}(q, \boldsymbol{\theta})$ can be decomposed into two different ways:

$$-\mathcal{F}(q, \boldsymbol{\theta}) = \ln p(\boldsymbol{x}|\boldsymbol{\theta}) - \mathrm{KL}[q(\boldsymbol{Z}) \| p(\boldsymbol{Z}|\boldsymbol{x}, \boldsymbol{\theta})], \qquad \textbf{(E step)}$$

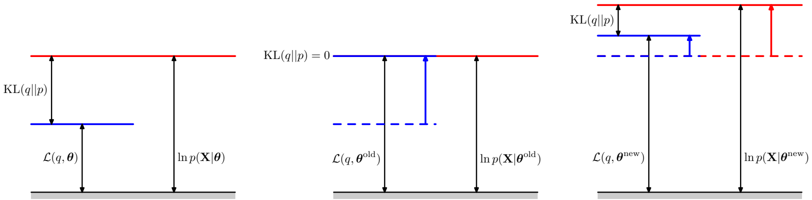
$$-\mathcal{F}(q, \boldsymbol{\theta}) = \langle \ln p(\boldsymbol{x}, \boldsymbol{Z}|\boldsymbol{\theta}) \rangle_{q(\boldsymbol{Z})} + \mathrm{H}[q(\boldsymbol{Z})]. \qquad \textbf{(M step)}$$

- EM maximises the lower bound by alternating between these two steps; it converges to local optimum of $\ln p(\boldsymbol{x}|\boldsymbol{\theta})$.
- By construction, the EM algorithm ensures a **monotonic** increase of the bound.
- Still ok if q is a good approximation of the true posterior (approximate E step).
- EM can be viewed as type II maximum likelihood (ML2).

EM in pictures



EM in pictures

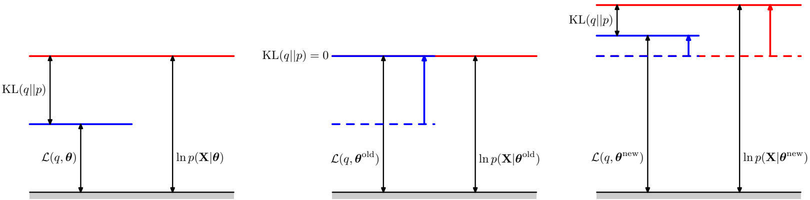


- Maximise lower bound by alternating between:

  E step: Set $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{x}, \boldsymbol{\theta})$ for fixed $\boldsymbol{\theta}$.

  M step: Maximise $\langle \ln p(\boldsymbol{x}, \boldsymbol{Z}|\boldsymbol{\theta}) \rangle$ for given $q(\boldsymbol{Z})$.

EM in pictures



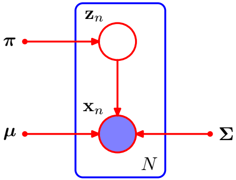
- Maximise lower bound by alternating between:

  E step: Set $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{x}, \boldsymbol{\theta})$ for fixed $\boldsymbol{\theta}$.

  M step: Maximise $\langle \ln p(\boldsymbol{x}, \boldsymbol{Z}|\boldsymbol{\theta}) \rangle$ for given $q(\boldsymbol{Z})$.

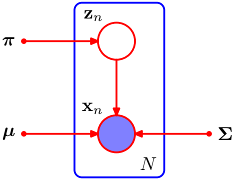
- Gradient ascent to local maxima of $\ln p(\boldsymbol{x}|\boldsymbol{\theta})$.

Mixture of Gaussians



$$p(\boldsymbol{x}|z) = \text{Gaussian}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z),$$
$$P(z) = \text{Categorical}\,(\boldsymbol{\pi})\,.$$

Mixture of Gaussians

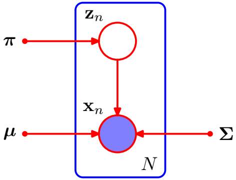


$$p(\boldsymbol{x}|z) = \text{Gaussian}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z),$$
$$P(z) = \text{Categorical}\,(\boldsymbol{\pi})\,.$$

- Log-complete likelihood:

$$\ln \prod_i p(\boldsymbol{x}_i, z_i) = \sum_i \sum_k \delta_k(z_i)\,(\ln \pi_k + \ln \text{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))\,.$$

Mixture of Gaussians



$$p(\boldsymbol{x}|z) = \mathrm{Gaussian}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z),$$
$$P(z) = \mathrm{Categorical}(\boldsymbol{\pi}).$$

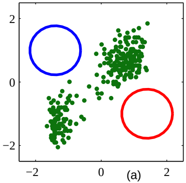
- Log-complete likelihood:

$$\ln \prod_i p(\boldsymbol{x}_i, z_i) = \sum_i \sum_k \delta_k(z_i) \left(\ln \pi_k + \ln \mathrm{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right).$$

- Responsibilities (E step):

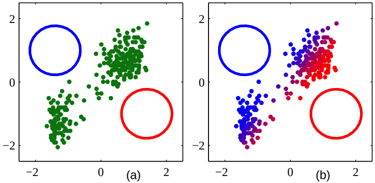
$$\rho_{ki} \equiv P(z = k|\boldsymbol{x}_i) = \frac{\pi_k \mathrm{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_l \pi_l \mathrm{Gaussian}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}. \qquad (\star)$$

Mixture of Gaussians (Old Faithful geyser data)

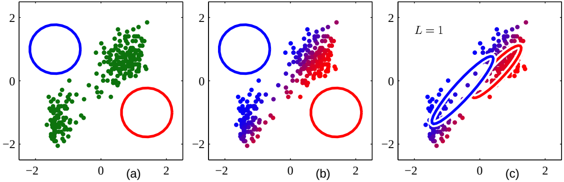


(a)

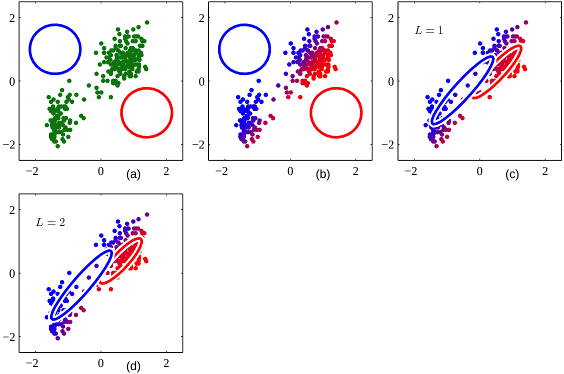
Mixture of Gaussians (Old Faithful geyser data)



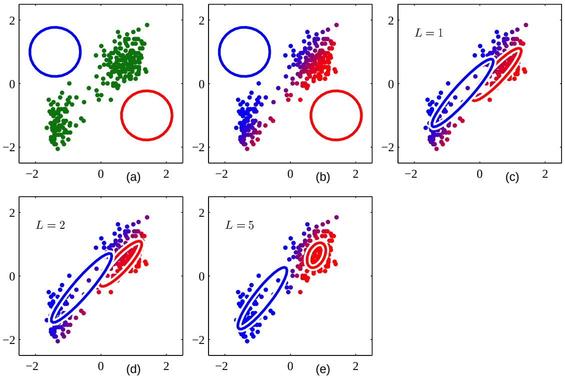
Mixture of Gaussians (Old Faithful geyser data)



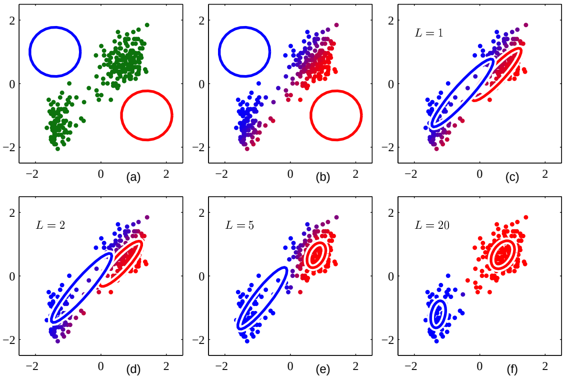
Mixture of Gaussians (Old Faithful geyser data)



Mixture of Gaussians (Old Faithful geyser data)



Mixture of Gaussians (Old Faithful geyser data)



Relation to Kmeans

1. Assign data point $\boldsymbol{x}_i$ to its closest cluster:

$$r_{ki} = \begin{cases} 1 & \text{if } k = \arg\min_l \|\boldsymbol{x}_i - \boldsymbol{\mu}_l\|^2, \\ 0 & \text{otherwise.} \end{cases}$$

2. Recompute the cluster means after having assigned all data points.

Relation to Kmeans

1. Assign data point $\boldsymbol{x}_i$ to its closest cluster:

$$r_{ki} = \begin{cases} 1 & \text{if } k = \arg\min_l \|\boldsymbol{x}_i - \boldsymbol{\mu}_l\|^2, \\ 0 & \text{otherwise.} \end{cases}$$

2. Recompute the cluster means after having assigned all data points.

Let us consider $p_{\boldsymbol{\theta}_k}(\boldsymbol{x}) = \text{Gaussian}(\boldsymbol{\mu}_k, \epsilon \boldsymbol{I})$:

$$\rho_{ki} = \frac{\pi_k \exp\left(-\frac{1}{2\epsilon}\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2\right)}{\sum_l \pi_l \exp\left(-\frac{1}{2\epsilon}\|\boldsymbol{x}_i - \boldsymbol{\mu}_l\|^2\right)}.$$

Relation to Kmeans

1. Assign data point $\boldsymbol{x}_i$ to its closest cluster:

$$r_{ki} = \begin{cases} 1 & \text{if } k = \arg\min_l \|\boldsymbol{x}_i - \boldsymbol{\mu}_l\|^2, \\ 0 & \text{otherwise.} \end{cases}$$

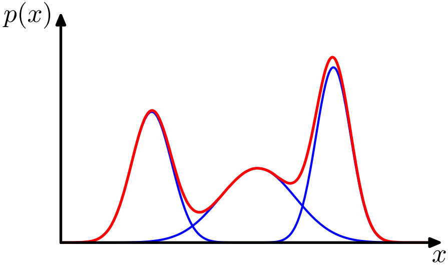
2. Recompute the cluster means after having assigned all data points.

Let us consider $p_{\boldsymbol{\theta}_k}(\boldsymbol{x}) = \mathrm{Gaussian}(\boldsymbol{\mu}_k, \epsilon \boldsymbol{I})$:

$$\lim_{\epsilon \to 0} \rho_{ki} = \lim_{\epsilon \to 0} \frac{\pi_k \exp\left(-\frac{1}{2\epsilon}\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2\right)}{\sum_l \pi_l \exp\left(-\frac{1}{2\epsilon}\|\boldsymbol{x}_i - \boldsymbol{\mu}_l\|^2\right)}.$$

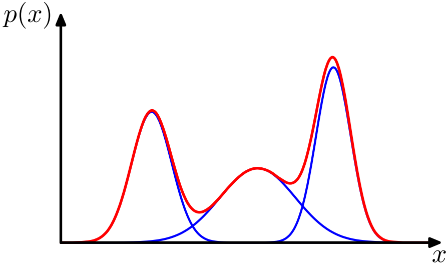
Other use cases?

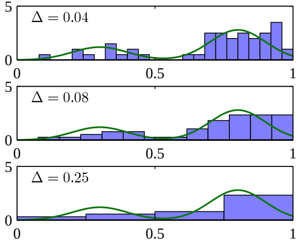
Density estimation:



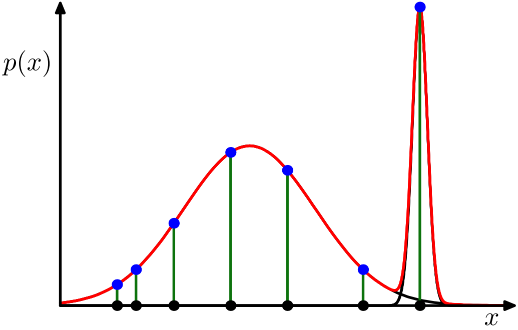
Other use cases?

Density estimation:



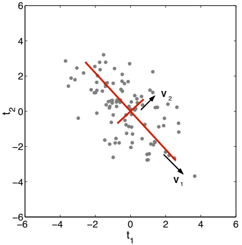


Failure mode



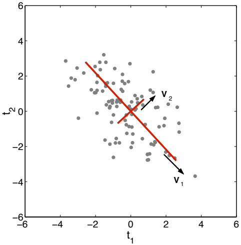
Probabilistic principal component analysis (PPCA)

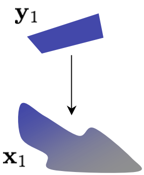
- PCA is a standard pre-processing tool for (linear) dimensionality reduction.
- It uses a maximal variance criterion (or minimal mean squared reconstruction error).
- Standard algorithms are $\mathcal{O}(D^3)$ (e.g. Gaussian elimination).



Probabilistic principal component analysis (PPCA)

- PCA is a standard pre-processing tool for (linear) dimensionality reduction.
- It uses a maximal variance criterion (or minimal mean squared reconstruction error).
- Standard algorithms are $\mathcal{O}(D^3)$ (e.g. Gaussian elimination).

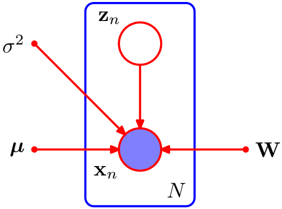




- PPCA assumes a single Gaussian latent variable and a Gaussian likelihood.
- ML solution spans same subspace as PCA solution.
- Standard EM is $\mathcal{O}(DNd)$ per iteration.

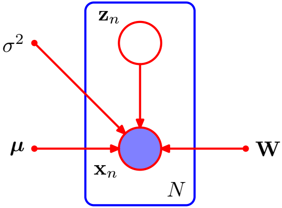
Probabilistic principal component analysis (PPCA)

$$\boldsymbol{x}_i = \boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$$



Probabilistic principal component analysis (PPCA)

$$\boldsymbol{x}_i = \boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$$

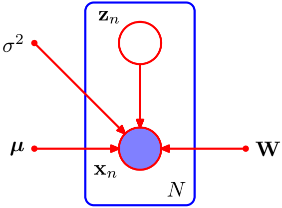


- Likelihood (noise model):

$$\boldsymbol{x}_i | \boldsymbol{z}_i \sim \mathrm{Gaussian}(\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_D).$$

Probabilistic principal component analysis (PPCA)

$$\boldsymbol{x}_i = \boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$$



- Likelihood (noise model):

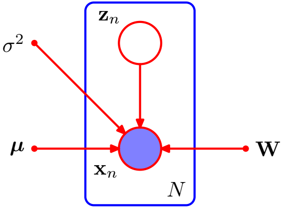
$$\boldsymbol{x}_i | \boldsymbol{z}_i \sim \mathrm{Gaussian}(\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_D).$$

- Continuous latent variable:

$$\boldsymbol{z}_i \sim \mathrm{Gaussian}(\mathbf{0}, \boldsymbol{I}_d).$$

Probabilistic principal component analysis (PPCA)

$$\boldsymbol{x}_i = \boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$$



- Likelihood (noise model):

$$\boldsymbol{x}_i | \boldsymbol{z}_i \sim \mathrm{Gaussian}(\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_D).$$

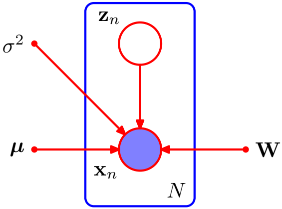
- Continuous latent variable:

$$\boldsymbol{z}_i \sim \mathrm{Gaussian}(\mathbf{0}, \boldsymbol{I}_d).$$

- ML estimate of the projection matrix: $\boldsymbol{W} = \boldsymbol{U}_d(\Lambda_d - \sigma^2 \boldsymbol{I}_d)^{1/2}\boldsymbol{R}$.

Probabilistic principal component analysis (PPCA)

$$\boldsymbol{x}_i = \boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$$



- Likelihood (noise model):

$$\boldsymbol{x}_i | \boldsymbol{z}_i \sim \text{Gaussian}(\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_D).$$

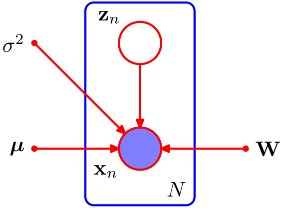
- Continuous latent variable:

$$\boldsymbol{z}_i \sim \text{Gaussian}(\boldsymbol{0}, \boldsymbol{I}_d).$$

- ML estimate of the projection matrix: $\boldsymbol{W} = \boldsymbol{U}_d(\Lambda_d - \sigma^2 \boldsymbol{I}_d)^{1/2}\boldsymbol{R}$.
- ML estimate is equivalent to PCA solution up to a rotation $\boldsymbol{R}$.

Probabilistic principal component analysis (PPCA)

$$\boldsymbol{x}_i = \boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$$



- Likelihood (noise model):

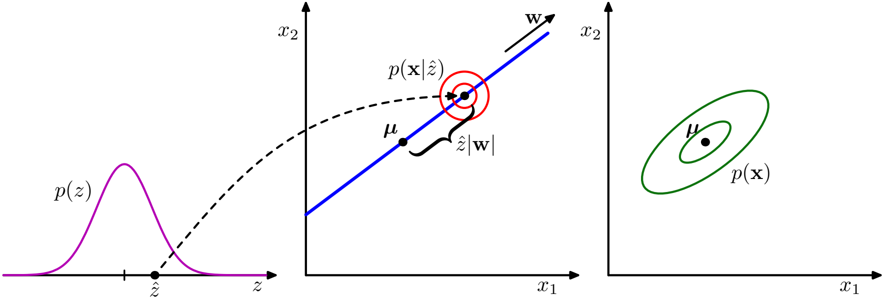
$$\boldsymbol{x}_i | \boldsymbol{z}_i \sim \mathrm{Gaussian}(\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_D).$$

- Continuous latent variable:

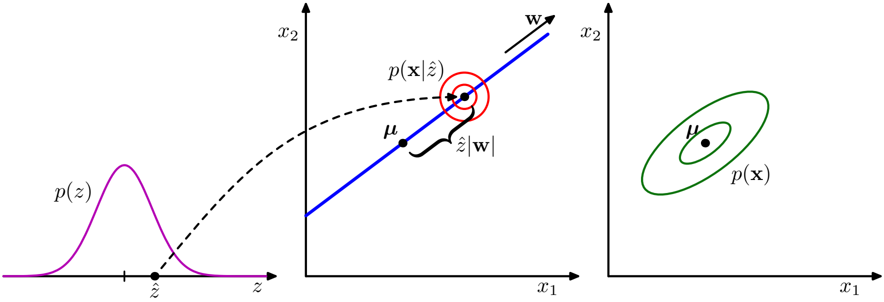
$$\boldsymbol{z}_i \sim \mathrm{Gaussian}(\mathbf{0}, \boldsymbol{I}_d).$$

- ML estimate of the projection matrix: $\boldsymbol{W} = \boldsymbol{U}_d(\Lambda_d - \sigma^2 \boldsymbol{I}_d)^{1/2}\boldsymbol{R}$.
- ML estimate is equivalent to PCA solution up to a rotation $\boldsymbol{R}$.
- Residual variance σ^2 is given by $\frac{1}{D-d}\sum_{j>d}\lambda_j$.

PPCA: interpretation



PPCA: interpretation



$$p(\boldsymbol{x}) = \text{Gaussian}(\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2 \boldsymbol{I}_D).$$

Mixtures of probabilistic principal component analysers

$$
\begin{aligned}
p(\boldsymbol{x}) &= \textstyle\sum_k \pi_k p(\boldsymbol{x}|z=k), \\
p(\boldsymbol{x}|z=k) &= \mathrm{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{W}_k \boldsymbol{W}_k^\top + \sigma^2 \boldsymbol{I}_D), \\
P(z) &= \mathrm{Categorical}\left(\boldsymbol{\pi}\right).
\end{aligned}
$$

Mixtures of probabilistic principal component analysers

$$
\begin{aligned}
p(\boldsymbol{x}) &= \textstyle\sum_k \pi_k p(\boldsymbol{x}|z=k), \\
p(\boldsymbol{x}|z=k) &= \text{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{W}_k \boldsymbol{W}_k^\top + \sigma^2 \boldsymbol{I}_D), \\
P(z) &= \text{Categorical}\,(\boldsymbol{\pi})\,.
\end{aligned}
$$

- Clustering (very) high-dimensional data:
  - Stable due to low rank approximation of the covariance matrices.
  - Captures correlations between local leading directions.
  - Rotational ambiguity vanishes.

Mixtures of probabilistic principal component analysers

$$
\begin{aligned}
p(\boldsymbol{x}) &= \textstyle\sum_k \pi_k p(\boldsymbol{x}|z=k), \\
p(\boldsymbol{x}|z=k) &= \mathrm{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{W}_k\boldsymbol{W}_k^\top + \sigma^2\boldsymbol{I}_D), \\
P(z) &= \mathrm{Categorical}\,(\boldsymbol{\pi})\,.
\end{aligned}
$$

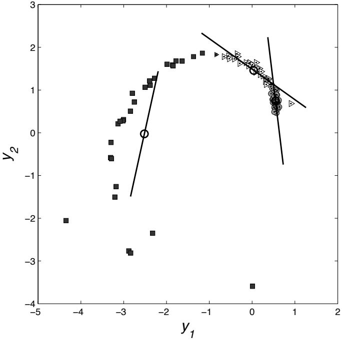
- Clustering (very) high-dimensional data:
  - Stable due to low rank approximation of the covariance matrices.
  - Captures correlations between local leading directions.
  - Rotational ambiguity vanishes.
- Combining local analysers to obtain nonlinear generative models.

Mixtures of probabilistic principal component analysers

$$
\begin{aligned}
p(\boldsymbol{x}) &= \textstyle\sum_k \pi_k p(\boldsymbol{x}|z=k), \\
p(\boldsymbol{x}|z=k) &= \mathrm{Gaussian}(\boldsymbol{\mu}_k, \boldsymbol{W}_k \boldsymbol{W}_k^\top + \sigma^2 \boldsymbol{I}_D), \\
P(z) &= \mathrm{Categorical}\left(\boldsymbol{\pi}\right).
\end{aligned}
$$

- Clustering (very) high-dimensional data:
  - Stable due to low rank approximation of the covariance matrices.
  - Captures correlations between local leading directions.
  - Rotational ambiguity vanishes.
- Combining local analysers to obtain nonlinear generative models.
- Possible issues are component misalignments and dimension mismatches.

Example



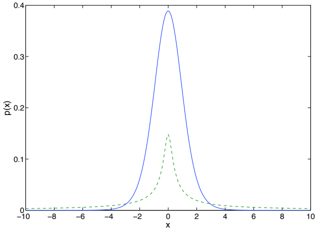
Can we fix this?

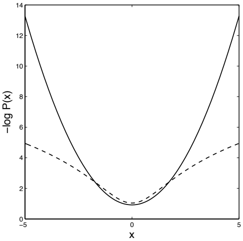
Can we fix this?

- Models based on Gaussian noise are sensitive to outliers!

Can we fix this?

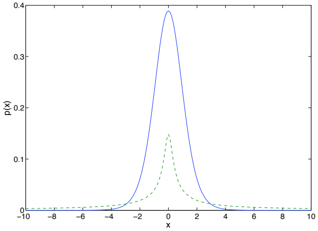
- Models based on Gaussian noise are sensitive to outliers!
- A robust reformulation is based on the Student-t density:

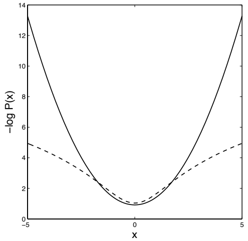




Can we fix this?

- Models based on Gaussian noise are sensitive to outliers!
- A robust reformulation is based on the Student-t density:





- Replace the Gaussian components by Student-t components:

$$
\begin{aligned}
p(\boldsymbol{x}) &= \textstyle\sum_k \pi_k p(\boldsymbol{x}|z=k), \\
p(\boldsymbol{x}|z=k) &= \mathrm{Student}(\boldsymbol{\mu}_k, \boldsymbol{W}_k\boldsymbol{W}_k^\top + \sigma^2 \boldsymbol{I}_D, \nu_k), \\
P(z) &= \mathrm{Categorical}\left(\boldsymbol{\pi}\right).
\end{aligned}
$$

Multivariate Student-t density

The Student-t density is defined as follows:[1]

$$\text{Student}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \left(1 + \frac{1}{\nu}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)^{-\frac{\nu+D}{2}}.$$

Parameter $\nu > 0$ is the shape parameter:

- The Cauchy density is recovered for $\nu = 1$.
- The Gaussian density is recovered when $\nu \to \infty$.

[1] Student's t density was published in 1908 by *William S. Gosset*, while he worked at Guinness Brewery in Dublin and was not allowed to publish under his own name.

Multivariate Student-t density

The Student-t density is defined as follows:[1]

$$\mathrm{Student}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \left(1 + \frac{1}{\nu}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)^{-\frac{\nu+D}{2}}.$$

Parameter $\nu > 0$ is the shape parameter:

- The Cauchy density is recovered for $\nu = 1$.
- The Gaussian density is recovered when $\nu \to \infty$.

The Student-t density can be reformulated as an infinite mixture of scaled Gaussians:

$$\mathrm{Student}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \int_0^\infty \mathrm{Gaussian}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/u)\ \mathrm{Gamma}(\tfrac{\nu}{2}, \tfrac{\nu}{2})\ du,$$

where u is a (latent) scale parameter.

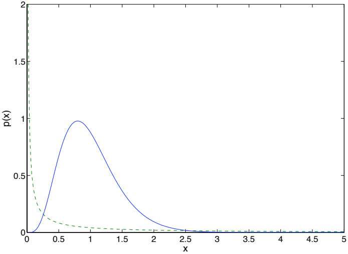
[1] Student's t density was published in 1908 by *William S. Gosset*, while he worked at Guinness Brewery in Dublin and was not allowed to publish under his own name.

Gamma density

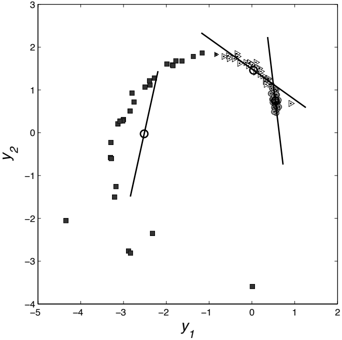
For $x \in \mathbb{R}^+$, the Gamma density is defined as follows:

$$\mathrm{Gamma}(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\}, \quad \alpha, \beta > 0,$$

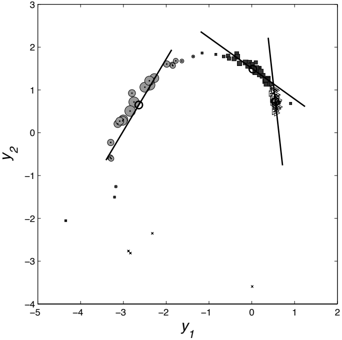
where $\Gamma(u) \equiv \int_0^\infty v^{u-1} e^{-v} dv$ is the *gamma function*.



Example (revisited)



(a) Standard PPCA.



(b) Robust PPCA.

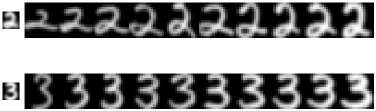
USPS handwritten digits 2 and 3

- USPS data set: 16×16 pixels images of digits (0 to 9).
- Only (respectively 731 and 658) images of digits 2 and 3 are kept.
- 100 (randomly chosen) images of digit 0.

USPS handwritten digits 2 and 3

- USPS data set: 16×16 pixels images of digits (0 to 9).
- Only (respectively 731 and 658) images of digits 2 and 3 are kept.
- 100 (randomly chosen) images of digit 0.

Mixture of PPCAs.



Mixture of robuts PPCAs.

USPS handwritten digits 2 and 3

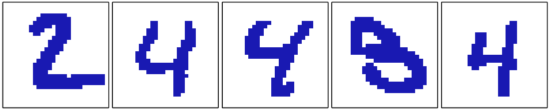
- USPS data set: 16×16 pixels images of digits (0 to 9).
- Only (respectively 731 and 658) images of digits 2 and 3 are kept.
- 100 (randomly chosen) images of digit 0.

Mixture of PPCAs.

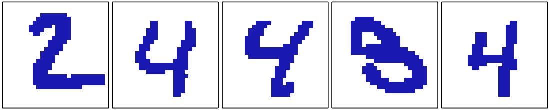
Mixture of robuts PPCAs.

Standard mixture of Gaussians and diagonal mixtures collapse...

Revisiting the digit recognition problem

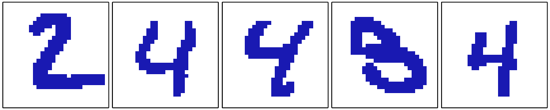


Revisiting the digit recognition problem



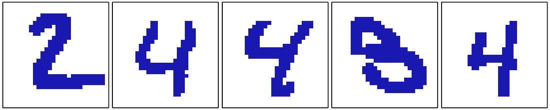
- Pixelised digits converted from grey scale to binary images by thresholding.

Revisiting the digit recognition problem



- Pixelised digits converted from grey scale to binary images by thresholding.
- Images are represented by a binary vector $\boldsymbol{x} = (x_1, \ldots, x_d)$.

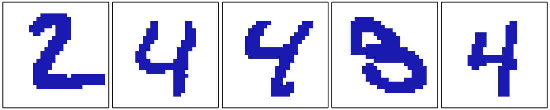
Revisiting the digit recognition problem



- Pixelised digits converted from grey scale to binary images by thresholding.
- Images are represented by a binary vector $\boldsymbol{x} = (x_1, \ldots, x_d)$.
- Goal is to cluster the images ($\sim$recognise digit automatically):

$$P(\boldsymbol{x}) = \sum_k \pi_k P_{\boldsymbol{\theta}_k}(\boldsymbol{x}), \qquad \sum_k \pi_k = 1, \qquad \pi_k \geqslant 0.$$

Revisiting the digit recognition problem



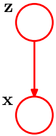
- Pixelised digits converted from grey scale to binary images by thresholding.
- Images are represented by a binary vector $\boldsymbol{x} = (x_1, \ldots, x_d)$.
- Goal is to cluster the images ($\sim$recognise digit automatically):

$$P(\boldsymbol{x}) = \sum_k \pi_k P_{\boldsymbol{\theta}_k}(\boldsymbol{x}), \qquad \sum_k \pi_k = 1, \qquad \pi_k \geqslant 0.$$

- Each component is a product of Bernoulli distributions:

$$P_{\boldsymbol{\theta}_k}(\boldsymbol{x}) = \prod_j \mathrm{Bernoulli}(\mu_{kj}).$$

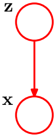
Mixture of Bernoulli distributions



$$P(\boldsymbol{x}|z) = \prod_j \mathrm{Bernoulli}(\mu_{zj}),$$

$$P(z) = \mathrm{Categorical}\left(\boldsymbol{\pi}\right).$$

Mixture of Bernoulli distributions



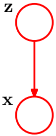
$$P(\boldsymbol{x}|z) = \prod_j \mathrm{Bernoulli}(\mu_{zj}),$$

$$P(z) = \mathrm{Categorical}\left(\boldsymbol{\pi}\right).$$

- Log-complete likelihood:

$$\ln \prod_i p(\boldsymbol{x}_i, z_i) = \sum_i \sum_k \delta_k(z_i) \left(\ln \pi_k + \sum_j \ln \mathrm{Bernoulli}(\mu_{zj}) \right).$$

Mixture of Bernoulli distributions



$$P(\boldsymbol{x}|z) = \prod_j \mathrm{Bernoulli}(\mu_{zj}),$$
$$P(z) = \mathrm{Categorical}(\boldsymbol{\pi}).$$

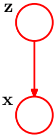
- Log-complete likelihood:

$$\ln \prod_i p(\boldsymbol{x}_i, z_i) = \sum_i \sum_k \delta_k(z_i) \left(\ln \pi_k + \sum_j \ln \mathrm{Bernoulli}(\mu_{zj}) \right).$$

- Responsibilities (E step):

$$\rho_{ki} \equiv P(z = k|\boldsymbol{x}_i) = \frac{\pi_k \prod_j \mathrm{Bernoulli}(\mu_{kj})}{\sum_l \pi_l \prod_j \mathrm{Bernoulli}(\mu_{lj})}.$$

Mixture of Bernoulli distributions



$$P(\boldsymbol{x}|z) = \prod_j \mathrm{Bernoulli}(\mu_{zj}),$$

$$P(z) = \mathrm{Categorical}(\boldsymbol{\pi}).$$

- Log-complete likelihood:

$$\ln \prod_i p(\boldsymbol{x}_i, z_i) = \sum_i \sum_k \delta_k(z_i) \left(\ln \pi_k + \sum_j \ln \mathrm{Bernoulli}(\mu_{zj}) \right).$$

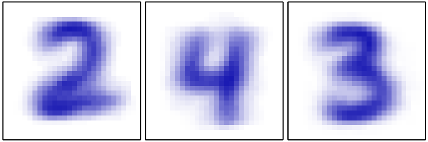
- Responsibilities (E step):

$$\rho_{ki} \equiv P(z = k|\boldsymbol{x}_i) = \frac{\pi_k \prod_j \mathrm{Bernoulli}(\mu_{kj})}{\sum_l \pi_l \prod_j \mathrm{Bernoulli}(\mu_{lj})}.$$

- Mean and mixture proportions (M step):

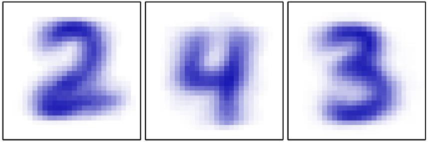
$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_i \rho_{ik} \boldsymbol{x}_i, \qquad \pi_k = \frac{n_k}{n}, \qquad n_k = \sum_i \rho_{ik}.$$

Cluster means

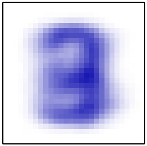


3 components

Cluster means



3 components



1 component

Outline

Admixtures

- Mixture model:

$$\begin{aligned} \boldsymbol{\pi} &\sim \text{Dirichlet}\left(\boldsymbol{\alpha}\right), \\ z_i | \boldsymbol{\pi} &\sim \text{Categorical}\left(\boldsymbol{\pi}\right), \\ \boldsymbol{x}_i | z_i &\sim p_{\boldsymbol{\theta}_{z_i}}(\boldsymbol{x}_i). \end{aligned}$$

Admixtures

- Mixture model:

$$\begin{aligned}\boldsymbol{\pi} &\sim \mathrm{Dirichlet}\left(\boldsymbol{\alpha}\right),\\ z_i|\boldsymbol{\pi} &\sim \mathrm{Categorical}\left(\boldsymbol{\pi}\right),\\ \boldsymbol{x}_i|z_i &\sim p_{\boldsymbol{\theta}_{z_i}}(\boldsymbol{x}_i).\end{aligned}$$

- Admixture model:

$$\begin{aligned}\boldsymbol{\pi}_i &\sim \mathrm{Dirichlet}\left(\boldsymbol{\alpha}\right),\\ z_{ij}|\boldsymbol{\pi}_i &\sim \mathrm{Categorical}\left(\boldsymbol{\pi}_i\right),\\ \boldsymbol{x}_{ij}|z_{ij} &\sim p_{\boldsymbol{\theta}_{z_{ij}}}(\boldsymbol{x}_{ij}).\end{aligned}$$

Dirichlet distribution

$$\boldsymbol{\mu} \sim \mathrm{Dirichlet}\,(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_j \mu_j^{\alpha_j - 1}, \qquad \alpha_j \geqslant 0.$$

Dirichlet distribution

$$\boldsymbol{\mu} \sim \mathrm{Dirichlet}\left(\boldsymbol{\alpha}\right) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_j \mu_j^{\alpha_j - 1}, \qquad \alpha_j \geqslant 0.$$

- Conjugate prior to the Multinomial distribution (and Categorical):

$$p(\boldsymbol{\mu}|\boldsymbol{x}) \propto P(\boldsymbol{x}|\boldsymbol{\mu})p(\boldsymbol{\mu}) \propto \prod_j \mu_j^{x_j + \alpha_j - 1}.$$

Dirichlet distribution

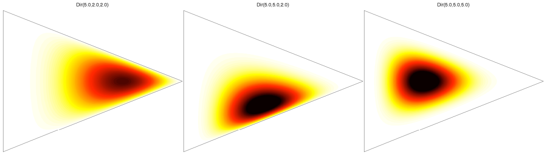
$$\boldsymbol{\mu} \sim \mathrm{Dirichlet}\,(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_j \mu_j^{\alpha_j - 1}, \qquad \alpha_j \geqslant 0.$$

- Conjugate prior to the Multinomial distribution (and Categorical):

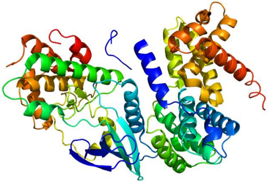
$$p(\boldsymbol{\mu}|\boldsymbol{x}) \propto P(\boldsymbol{x}|\boldsymbol{\mu})p(\boldsymbol{\mu}) \propto \prod_j \mu_j^{x_j + \alpha_j - 1}.$$

- Defines a distribution over the simplex:

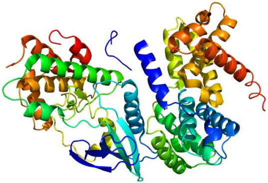
$$\sum_j \mu_j = 1, \qquad \mu_j \geqslant 0.$$



Topic models

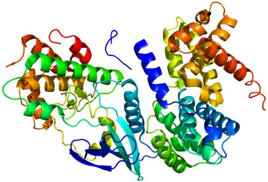


Topic models



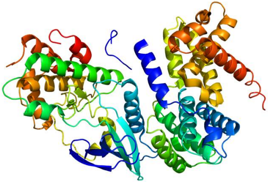
- Extremely popular (e.g., more than 14k citations in Google Scholar)

Topic models



- Extremely popular (e.g., more than 14k citations in Google Scholar)
- Organise and browse large document collections

Topic models



- Extremely popular (e.g., more than 14k citations in Google Scholar)
- Organise and browse large document collections
- Capture underlying semantic structure (in an unsupervised way)

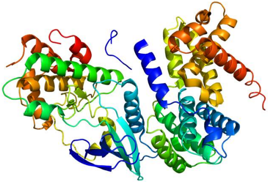
Topic models

stackoverflow

WIRED

IdeaScale where ideas come to life

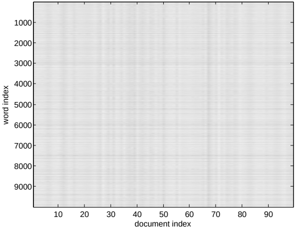
flickr



- Extremely popular (e.g., more than 14k citations in Google Scholar)
- Organise and browse large document collections
- Capture underlying semantic structure (in an unsupervised way)
- Easily extended to discover trends, to account for the author, to model multilingual documents, to relate to the social network, etc.

Latent Dirichlet allocation (LDA)

(Blei et al., JMLR 2003)



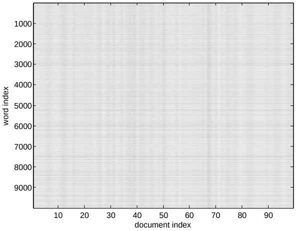
Observations are word counts per document. LDA assumes an admixture model:

$$\mathbf{X} \in \mathbb{N}^{V \times D},$$

$$\mathbf{x}_d \sim \prod_{i=1}^{N_d} \sum_k \theta_{kd} \mathrm{Categorical}(\phi_k).$$

Latent Dirichlet allocation (LDA)

(Blei et al., JMLR 2003)



Observations are word counts per document. LDA assumes an admixture model:

$$\mathbf{X} \in \mathbb{N}^{V \times D},$$

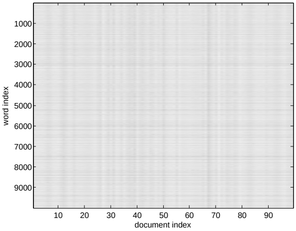
$$\mathbf{x}_d \sim \prod_{i=1}^{N_d} \sum_k \theta_{kd} \mathrm{Categorical}(\phi_k).$$

LDA infers a low-rank approximation of the matrix of counts:

$$\mathrm{E}(\mathbf{X}) \approx \boldsymbol{\Phi}\boldsymbol{\Theta}^\top, \qquad \mathbf{x}_d \sim \mathrm{Multinomial}(\boldsymbol{\Phi}\boldsymbol{\theta}_d, N_d)$$

where $\boldsymbol{\Phi} \in \mathbb{R}_+^{V \times K}$, $\boldsymbol{\Theta} \in \mathbb{R}_+^{D \times K}$ and K is small.

Latent Dirichlet allocation (LDA) (Blei et al., JMLR 2003)



Observations are word counts per document. LDA assumes an admixture model:

$$\mathbf{X} \in \mathbb{N}^{V \times D},$$

$$\mathbf{x}_d \sim \prod_{i=1}^{N_d} \sum_k \theta_{kd} \mathrm{Categorical}(\phi_k).$$

LDA infers a low-rank approximation of the matrix of counts:

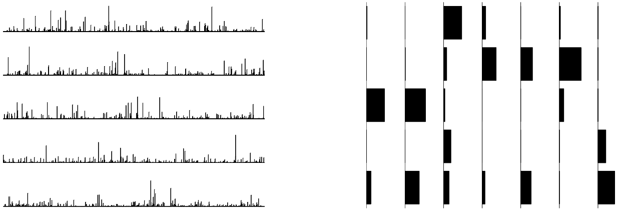
$$\mathrm{E}(\mathbf{X}) \approx \mathbf{\Phi}\mathbf{\Theta}^\top, \qquad \mathbf{x}_d \sim \mathrm{Multinomial}(\mathbf{\Phi}\boldsymbol{\theta}_d, N_d)$$

where $\mathbf{\Phi} \in \mathbb{R}_+^{V \times K}$, $\mathbf{\Theta} \in \mathbb{R}_+^{D \times K}$ and K is small.

Simple generative model for text, based on a bag-of-words representation.

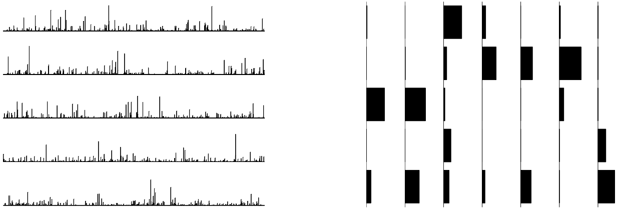
Generative model for documents

- Let V be the size of the vocabulary and K the number of topics.
- Topic k is defined as the categorical distribution ϕ_k over the vocabulary.
- Document d is summarised as a mixture of these topics.



Generative model for documents

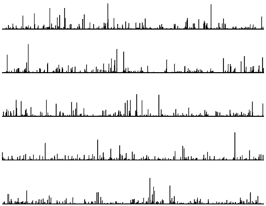
- Let V be the size of the vocabulary and K the number of topics.
- Topic k is defined as the categorical distribution ϕ_k over the vocabulary.
- Document d is summarised as a mixture of these topics.

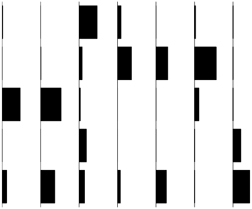


Document d is generated as follows:

Generative model for documents

- Let V be the size of the vocabulary and K the number of topics.
- Topic k is defined as the categorical distribution ϕ_k over the vocabulary.
- Document d is summarised as a mixture of these topics.



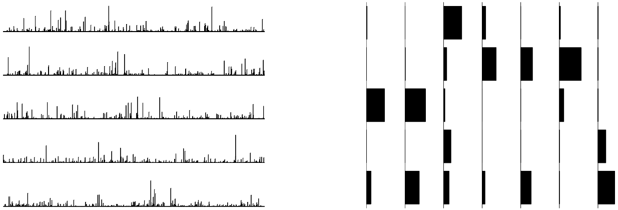


Document d is generated as follows:

1. The number of words N_d in document d is drawn from a Poisson.

Generative model for documents

- Let V be the size of the vocabulary and K the number of topics.
- Topic k is defined as the categorical distribution ϕ_k over the vocabulary.
- Document d is summarised as a mixture of these topics.

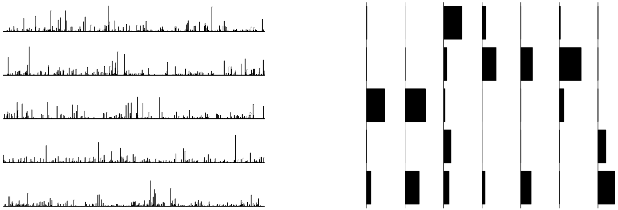


Document d is generated as follows:

1. The number of words N_d in document d is drawn from a Poisson.
2. The topic proportions $\boldsymbol{\theta}_d$ in document d are drawn from a Dirichlet; this vector defines a categorical distribution over the topics.

Generative model for documents

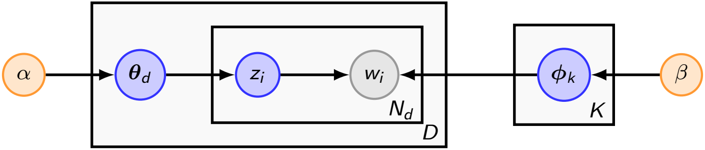
- Let V be the size of the vocabulary and K the number of topics.
- Topic k is defined as the categorical distribution ϕ_k over the vocabulary.
- Document d is summarised as a mixture of these topics.



Document d is generated as follows:

1. The number of words N_d in document d is drawn from a Poisson.
2. The topic proportions $\boldsymbol{\theta}_d$ in document d are drawn from a Dirichlet; this vector defines a categorical distribution over the topics.
3. The topic z_i associated to word w_i is drawn from $\boldsymbol{\theta}_d$; word w_i is then drawn from the categorical distribution ϕ_{z_i}.

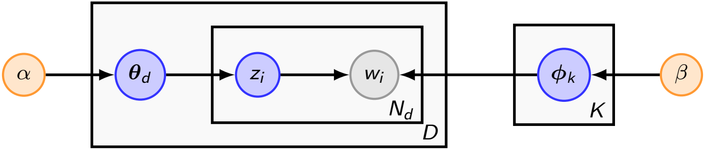
Graphical model and inference



$$\boldsymbol{\theta}_d \sim \text{Dirichlet}\left(\alpha \mathbf{1}_K\right), \qquad z_i | \boldsymbol{\theta}_d \sim \text{Categorical}(\boldsymbol{\theta}_d),$$

$$\boldsymbol{\phi}_k \sim \text{Dirichlet}\left(\beta \mathbf{1}_V\right), \qquad w_i | z_i, \{\boldsymbol{\phi}_k\}_{k=1}^{K} \sim \text{Categorical}(\boldsymbol{\phi}_{z_i}).$$

Graphical model and inference



$$\boldsymbol{\theta}_d \sim \text{Dirichlet}\left(\alpha \mathbf{1}_K\right), \qquad z_i | \boldsymbol{\theta}_d \sim \text{Categorical}(\boldsymbol{\theta}_d),$$

$$\boldsymbol{\phi}_k \sim \text{Dirichlet}\left(\beta \mathbf{1}_V\right), \qquad w_i | z_i, \{\boldsymbol{\phi}_k\}_{k=1}^K \sim \text{Categorical}(\boldsymbol{\phi}_{z_i}).$$

Collapsed Gibbs sampler (Griffiths and Steyvers, PNAS 2004):

$$p(z_i = k | \boldsymbol{w}, \boldsymbol{z}^{\setminus i}) \propto p(\boldsymbol{w}|\boldsymbol{z})p(\boldsymbol{z}) \propto \frac{(\alpha + n_{\cdot kd}^{\setminus i})(\beta + n_{vk\cdot}^{\setminus i})}{V\beta + n_{\cdot k \cdot}^{\setminus i}},$$

where n_{vkd} is the number of times word v is assigned to topic k in document d.

Applications and extensions of topic models

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Applications and extensions of topic models

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

- Author topic model

Applications and extensions of topic models

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

- Author topic model
- Topics over time

Applications and extensions of topic models

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

- Author topic model
- Topics over time
- N-gram topic models

Applications and extensions of topic models

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

- Author topic model
- Topics over time
- N-gram topic models
- Hierarchical topic models

Applications and extensions of topic models

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

- Author topic model
- Topics over time
- N-gram topic models
- Hierarchical topic models
- Multi-lingual topic models
- Topic model for images
- Population genetics
- ...

Outline

Summary

- Gaussian, Student, Bernoulli mixtures
- Alternative view of EM algorithm
- Latent Dirichlet Allocation

Outline

Exercise

Derive the M step for a mixture of Gaussians.

References

C. Archambeau, et al. (2008): *Mixtures of Robust Probabilistic Principal Component Analyzers*. Neurocomputing, 71(7-9):1274-1282, 2008.

C. Bishop: *Pattern Recognition and Machine Learning*. Springer, 2006.

D. Blei, et al. (2003): *Latent Dirichlet Allocation*. Journal of Machine Learning Research 3:993-1022.

T. L. Griffiths and M. Steyvers (2003): *Finding scientific topics*. Proceedings of the PNAS.

R. M. Neal and G. Hinton (1998): *A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants*.