Dimension Reduction (PCA, ICA, CCA, FLD, Topic Models)

Yi Zhang

10-701, Machine Learning, Spring 2011

April 6^{th}, 2011

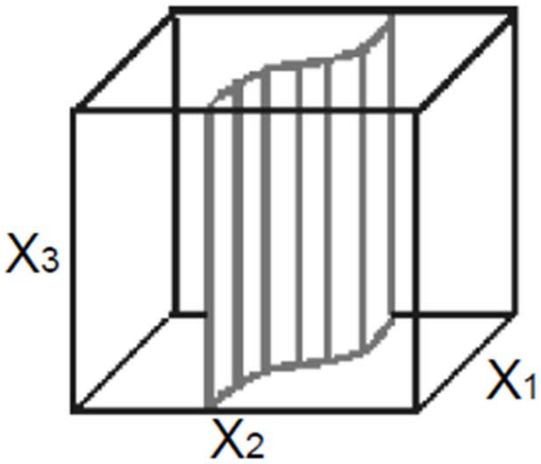
Parts of the PCA slides are from previous 10-701 lectures

Outline

- **Dimension reduction**
- Principal Components Analysis
- Independent Component Analysis
- Canonical Correlation Analysis
- Fisher's Linear Discriminant
- Topic Models and Latent Dirichlet Allocation

Dimension reduction

- Feature selection – select a subset of features

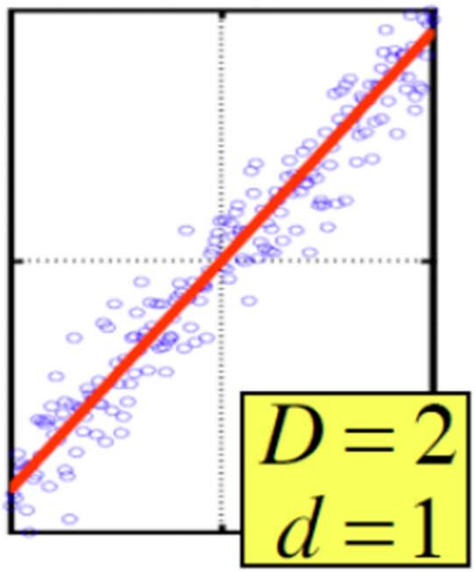


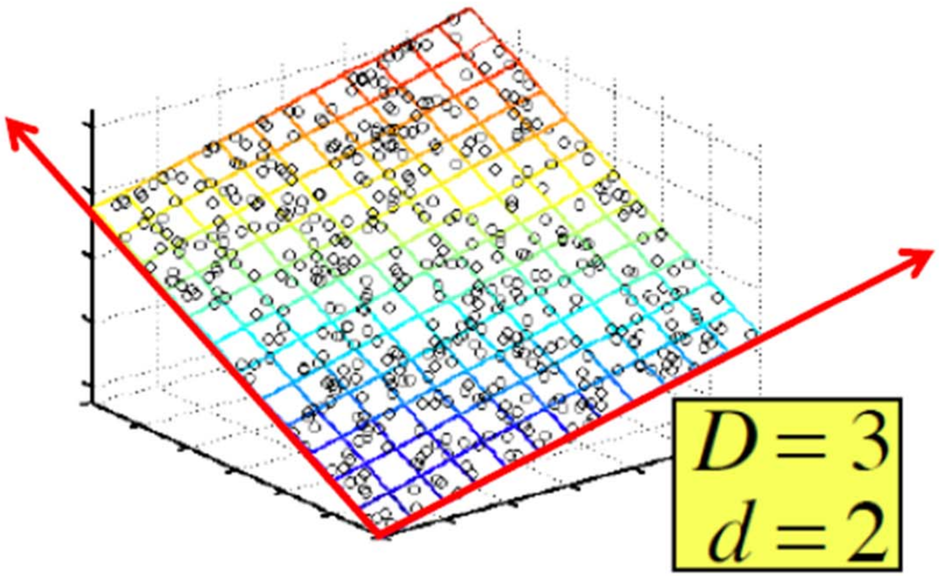
X3 - Irrelevant

- More generally, ***feature extraction***
  - Not limited to the original features
  - "Dimension reduction" usually refers to this case

Dimension reduction

- Assumption: data (approximately) lies on a lower dimensional space
- Examples:

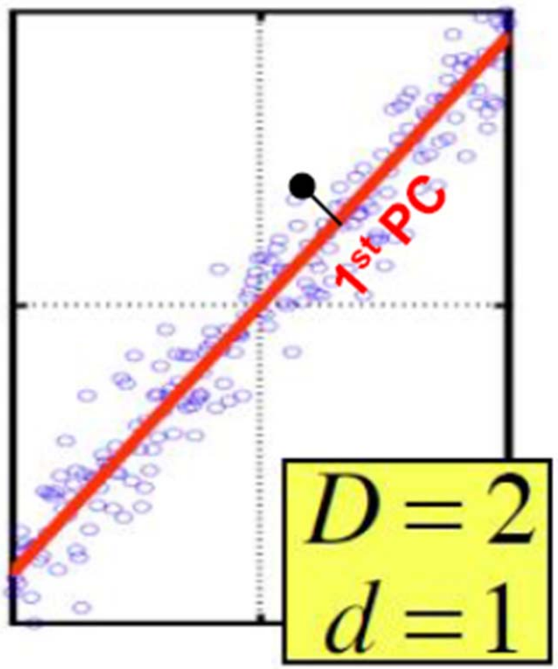




Outline

- Dimension reduction
- **Principal Components Analysis**
- Independent Component Analysis
- Canonical Correlation Analysis
- Fisher's Linear Discriminant
- Topic Models and Latent Dirichlet Allocation

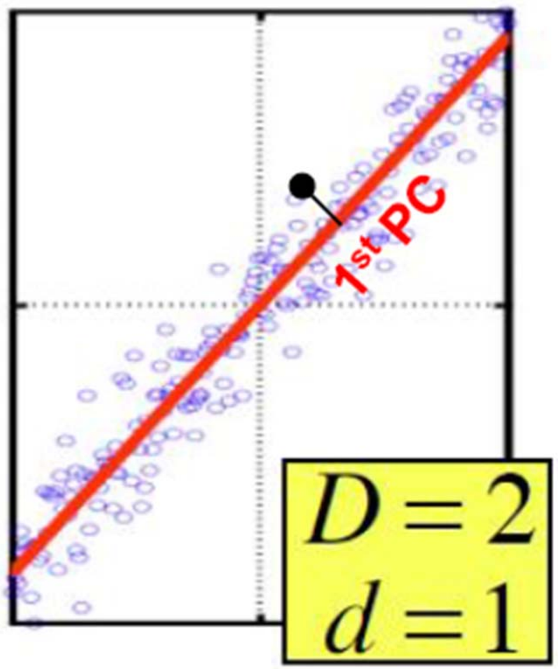
Principal components analysis



Principal Components (PC) are orthogonal directions that capture most of the variance in the data

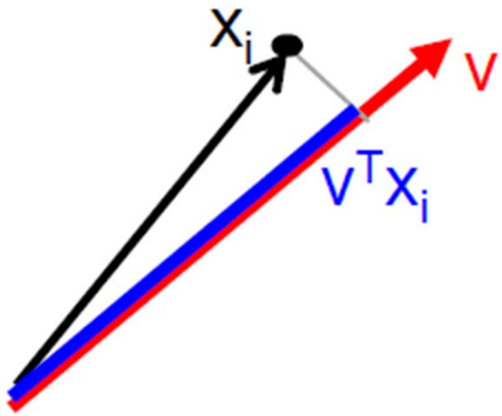
1st PC – direction of greatest variability in data

Principal components analysis



Principal Components (PC) are orthogonal directions that capture most of the variance in the data

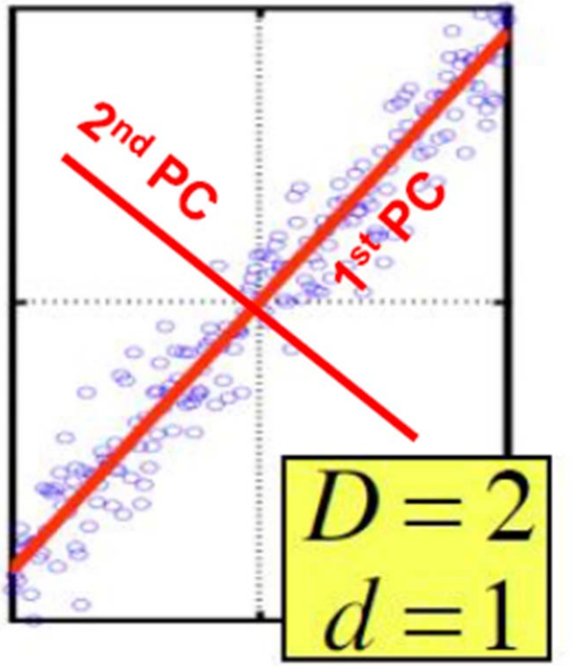
1st PC – direction of greatest variability in data



Take a data point x_i (D-dimensional vector)

Projection of x_i onto the 1st PC v is $v^T x_i$

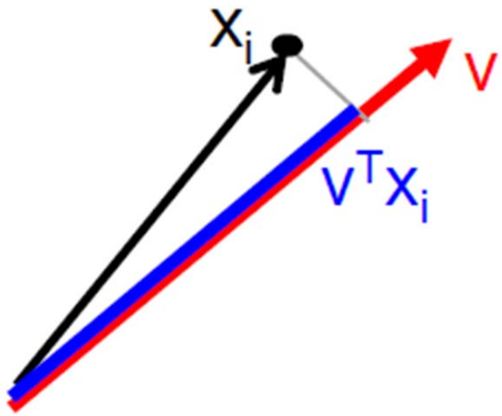
Principal components analysis



Principal Components (PC) are orthogonal directions that capture most of the variance in the data

1st PC – direction of greatest variability in data

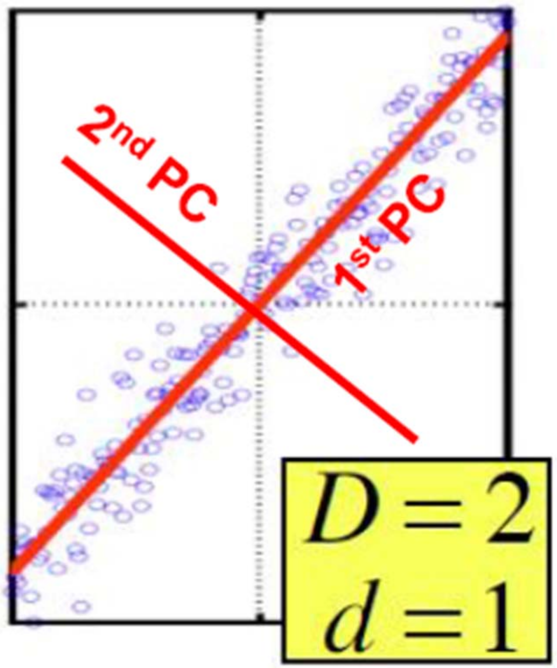
2nd PC – Next orthogonal (uncorrelated) direction of greatest variability



Take a data point x_i (D-dimensional vector)

Projection of x_i onto the 1st PC v is $v^T x_i$

Principal components analysis

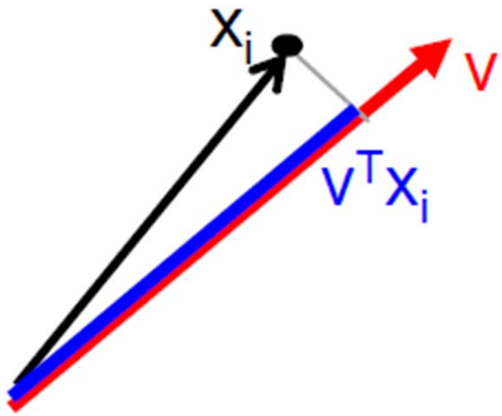


Principal Components (PC) are orthogonal directions that capture most of the variance in the data

1st PC – direction of greatest variability in data

2nd PC – Next orthogonal (uncorrelated) direction of greatest variability

(remove all variability in first direction, then find next direction of greatest variability)



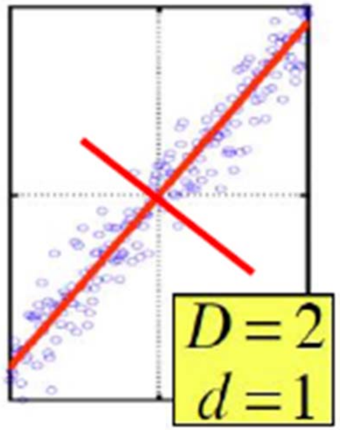
Take a data point x_i (D-dimensional vector)

Projection of x_i onto the 1st PC v is $v^T x_i$

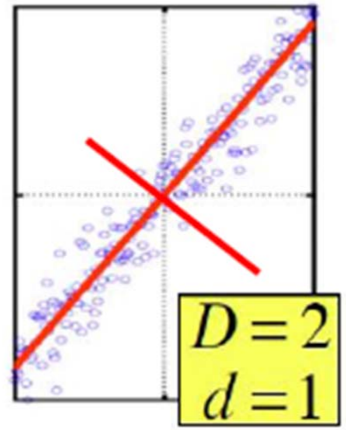
Principal components analysis

- Assume data is centered
- For a projection direction v
  - Variance of projected data

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v}$$



Principal components analysis



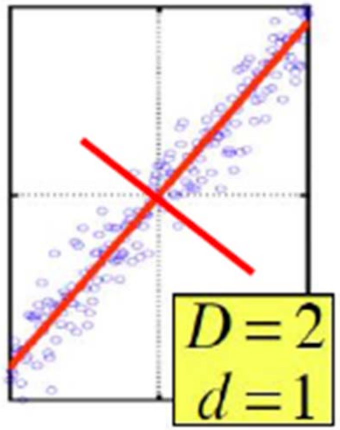
- Assume data is centered
- For a projection direction v
  - Variance of projected data

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v}$$

  - Maximize the variance of projected data

$$\max_{\mathbf{v}} \ \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} \quad \text{s.t.} \quad \mathbf{v}^T\mathbf{v} = 1$$

Principal components analysis



- Assume data is centered
- For a projection direction v
  - Variance of projected data

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v}$$

  - Maximize the variance of projected data

$$\max_{\mathbf{v}} \ \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} \quad \text{s.t.} \quad \mathbf{v}^T\mathbf{v} = 1$$

  - How to solve this ?

Principal components analysis

- PCA formulation

$$\max_{\mathbf{v}} \ \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} \quad \text{s.t.} \quad \mathbf{v}^T\mathbf{v} = 1$$

Lagrangian: $\max_{\mathbf{v}} \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} - \lambda\mathbf{v}^T\mathbf{v}$

Wrap constraints into the objective function

$$\partial/\partial\mathbf{v} = 0 \qquad (\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I})\mathbf{v} = 0 \qquad \Rightarrow (\mathbf{X}\mathbf{X}^T)\mathbf{v} = \lambda\mathbf{v}$$

Therefore, v is the eigenvector of sample correlation/ covariance matrix XX^T

- As a result …

The 1st Principal component v_1 is the eigenvector of the sample covariance matrix XX^T associated with the largest eigenvalue λ_1

The 2nd Principal component v_2 is the eigenvector of the sample covariance matrix XX^T associated with the second largest eigenvalue λ_2

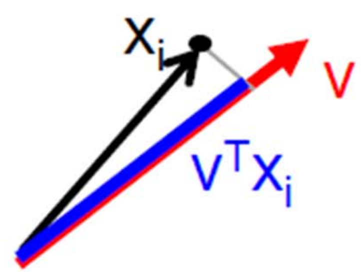
Principal components analysis

Maximum Variance Subspace: PCA finds vectors v such that projections on to the vectors capture maximum variance in the data

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} \qquad \text{s.t.} \qquad \mathbf{v}^T\mathbf{v} = 1$$

Minimum Reconstruction Error: PCA finds vectors v such that projection on to the vectors yields minimum MSE reconstruction

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - (\mathbf{v}^T\mathbf{x}_i)\mathbf{v}\|^2 \quad \text{s.t.} \qquad \mathbf{v}^T\mathbf{v} = 1$$

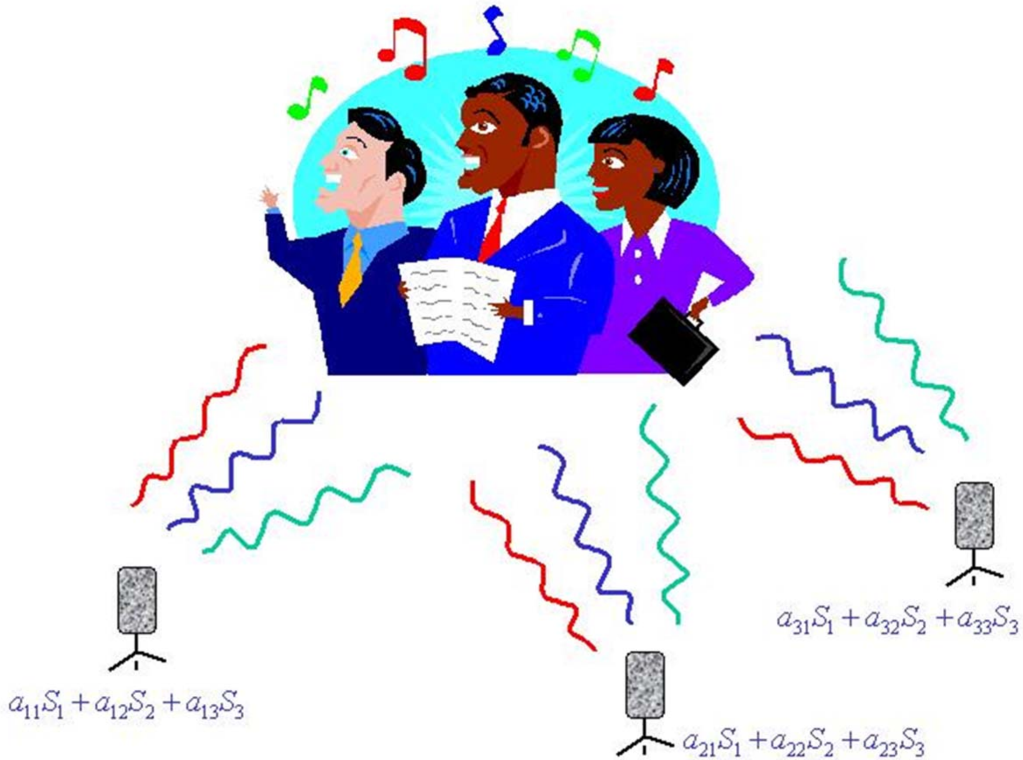


Outline

- Dimension reduction
- Principal Components Analysis
- **Independent Component Analysis**
- Canonical Correlation Analysis
- Fisher's Linear Discriminant
- Topic Models and Latent Dirichlet Allocation

Source separation

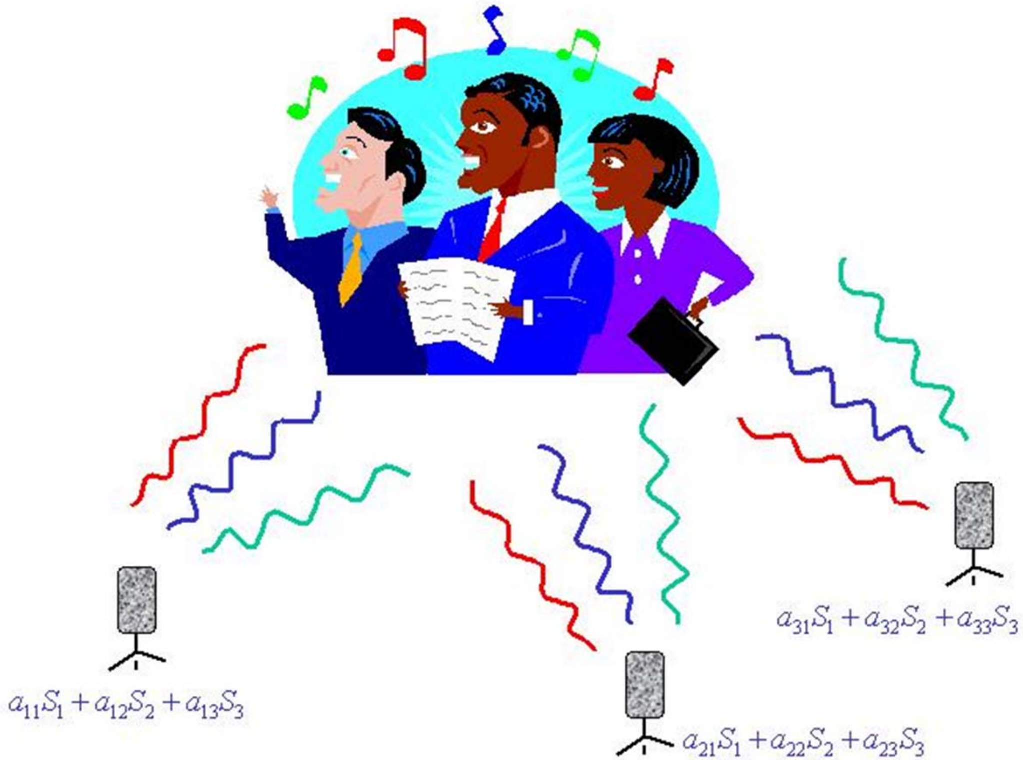
- The classical "cocktail party" problem



  - Separate the mixed signal into sources

Source separation

- The classical "cocktail party" problem



  - Separate the mixed signal into sources
  - Assumption: different sources are ***independent***

Independent component analysis

- Let $v_1, v_2, v_3, \ldots v_d$ denote the projection directions of independent components
- ICA: find these directions such that data projected onto these directions have maximum statistical independence

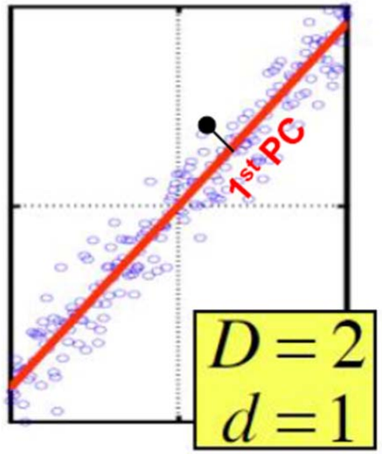
Independent component analysis

- Let $v_1, v_2, v_3, \ldots v_d$ denote the projection directions of independent components
- ICA: find these directions such that data projected onto these directions have maximum statistical independence
- How to actually maximize independence?
  - Minimize the mutual information
  - Or maximize the non-Gaussianity
  - Actual formulation quite complicated !

Outline

- Dimension reduction
- Principal Components Analysis
- Independent Component Analysis
- **Canonical Correlation Analysis**
- Fisher's Linear Discriminant
- Topic Models and Latent Dirichlet Allocation

Recall: PCA



- Principal component analysis

$$\max_{\mathbf{v}} \ \mathbf{v}^T \mathbf{X} \mathbf{X}^T \mathbf{v} \quad \text{s.t.} \quad \mathbf{v}^T \mathbf{v} = 1$$

  - Note: $\frac{1}{n} \sum_{i=1}^{n} (\mathbf{v}^T \mathbf{x}_i)^2 = \mathbf{v}^T \mathbf{X} \mathbf{X}^T \mathbf{v}$
  - Find the projection direction v such that the variance of projected data is maximized
  - Intuitively, find the intrinsic subspace of the original feature space (in terms of retaining the data variability)

Canonical correlation analysis

- Now consider ***two*** sets of variables **x** and **y**
  - **x** is a vector of *p* variables
  - **y** is a vector of *q* variables
  - Basically, ***two*** feature spaces
- How to find the connection between two set of variables (or two feature spaces)?

Canonical correlation analysis

- Now consider ***two*** sets of variables **x** and **y**
  - **x** is a vector of *p* variables
  - **y** is a vector of *q* variables
  - Basically, ***two*** feature spaces
- How to find the connection between two set of variables (or two feature spaces)?
  - CCA: find a projection direction **u** in the space of **x**, and a projection direction **v** in the space of **y**, so that projected data onto **u** and **v** has ***max correlation***
  - Note: CCA simultaneously finds dimension reduction for two feature spaces

Canonical correlation analysis

- CCA formulation

$$\underset{\mathbf{u}\in R^p, \mathbf{v}\in R^q}{\operatorname{argmax}} \quad \frac{\mathbf{u}^T \mathbf{X}^T \mathbf{Y} \mathbf{v}}{\sqrt{(\mathbf{u}^T \mathbf{X}^T \mathbf{X} \mathbf{u})(\mathbf{v}^T \mathbf{Y}^T \mathbf{Y} \mathbf{v})}}$$

  - **X** is n by p: n samples in p-dimensional space
  - **Y** is n by q: n samples in q-dimensional space
  - The n samples are ***paired*** in **X** and **Y**

Canonical correlation analysis

- CCA formulation

$$\underset{\mathbf{u}\in R^p,\mathbf{v}\in R^q}{\operatorname{argmax}} \quad \frac{\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v}}{\sqrt{(\mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u})(\mathbf{v}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v})}}$$

  - **X** is n by p: n samples in p-dimensional space
  - **Y** is n by q: n samples in q-dimensional space
  - The n samples are ***paired*** in **X** and **Y**
- How to solve? … kind of complicated …

Canonical correlation analysis

- CCA formulation

$$\underset{\mathbf{u}\in R^p,\mathbf{v}\in R^q}{\operatorname{argmax}} \quad \frac{\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v}}{\sqrt{(\mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u})(\mathbf{v}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v})}}$$

  - **X** is n by p: n samples in p-dimensional space
  - **Y** is n by q: n samples in q-dimensional space
  - The n samples are ***paired*** in **X** and **Y**
- How to solve? Generalized eigenproblems !

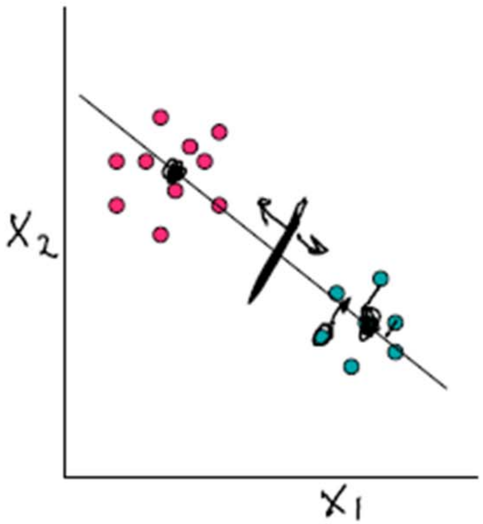
$$\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}\mathbf{u} = \lambda\mathbf{X}^T\mathbf{X}\mathbf{u}$$
$$\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{v} = \lambda\mathbf{Y}^T\mathbf{Y}\mathbf{v}$$

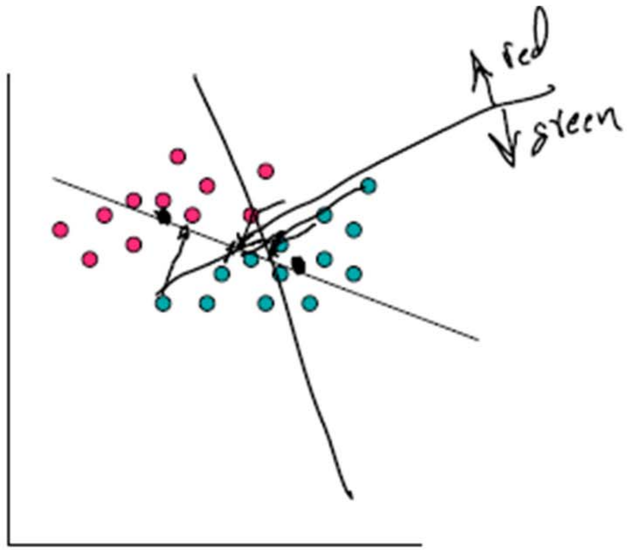
Outline

- Dimension reduction
- Principal Components Analysis
- Independent Component Analysis
- Canonical Correlation Analysis
- **Fisher's Linear Discriminant**
- Topic Models and Latent Dirichlet Allocation

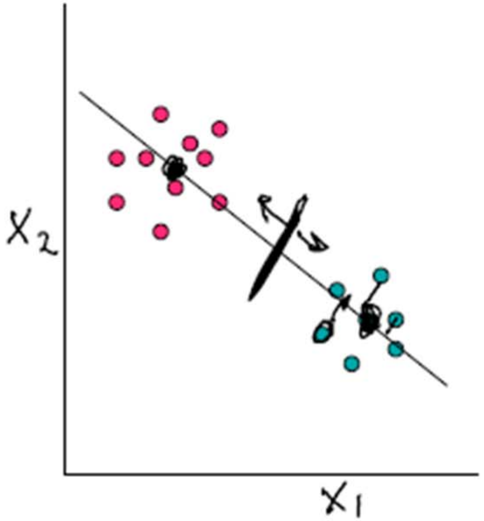
Fisher's linear discriminant

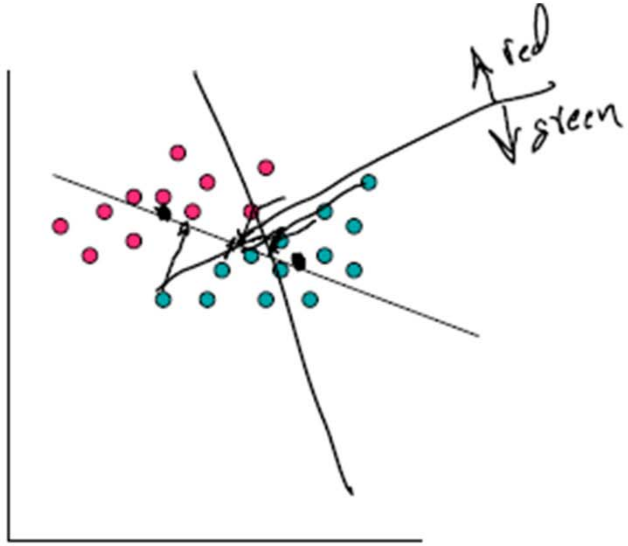
- Now come back to *one* feature space
- In addition to features, we also have ***label***
  - Find the dimension reduction that helps separate different classes of examples !
  - Let's consider 2-class case





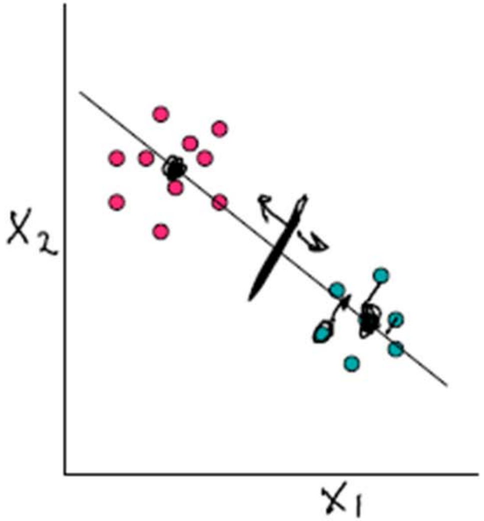
Fisher's linear discriminant

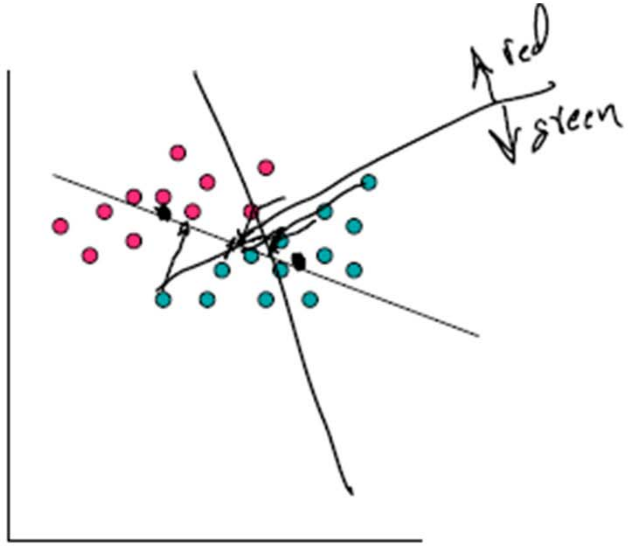




- Idea: maximize the ratio of "between-class variance" over "within-class variance" for the projected data

Fisher's linear discriminant





Fisher Linear Discriminant chooses: $\arg\max_{\mathbf{w}} \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$

$$m_i \equiv \mathbf{w}^T \mathbf{m}_i \qquad s_i^2 \equiv \sum_{n \in C_i} (x^n - m_i)^2$$

Fisher's linear discriminant

- Generalize to multi-class cases
- Still, maximizing the ratio of "between-class variance" over "within-class variance" of the projected data:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

$$\begin{aligned} S_B &= \sum_c (\boldsymbol{\mu}_c - \bar{\mathbf{x}})(\boldsymbol{\mu}_c - \bar{\mathbf{x}})^T \\ S_W &= \sum_c \sum_{i \in c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T \end{aligned}$$

Outline

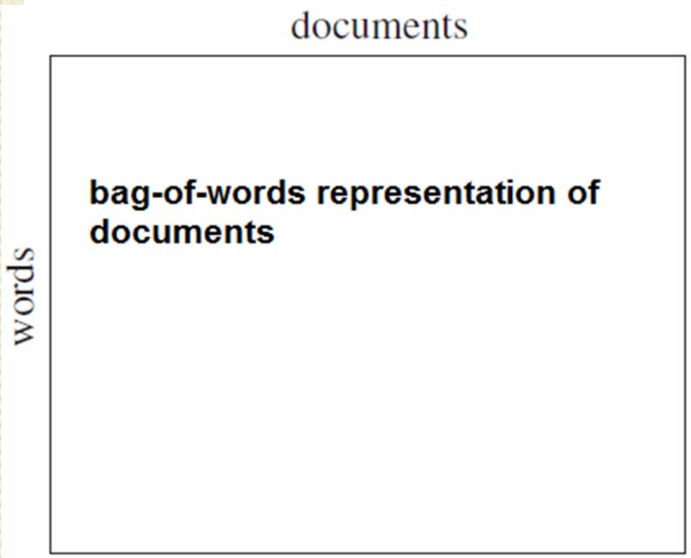
- Dimension reduction
- Principal Components Analysis
- Independent Component Analysis
- Canonical Correlation Analysis
- Fisher’s Linear Discriminant
- **Topic Models and Latent Dirichlet Allocation**

Topic models

- Topic models: a class of dimension reduction models on text (from words to topics)

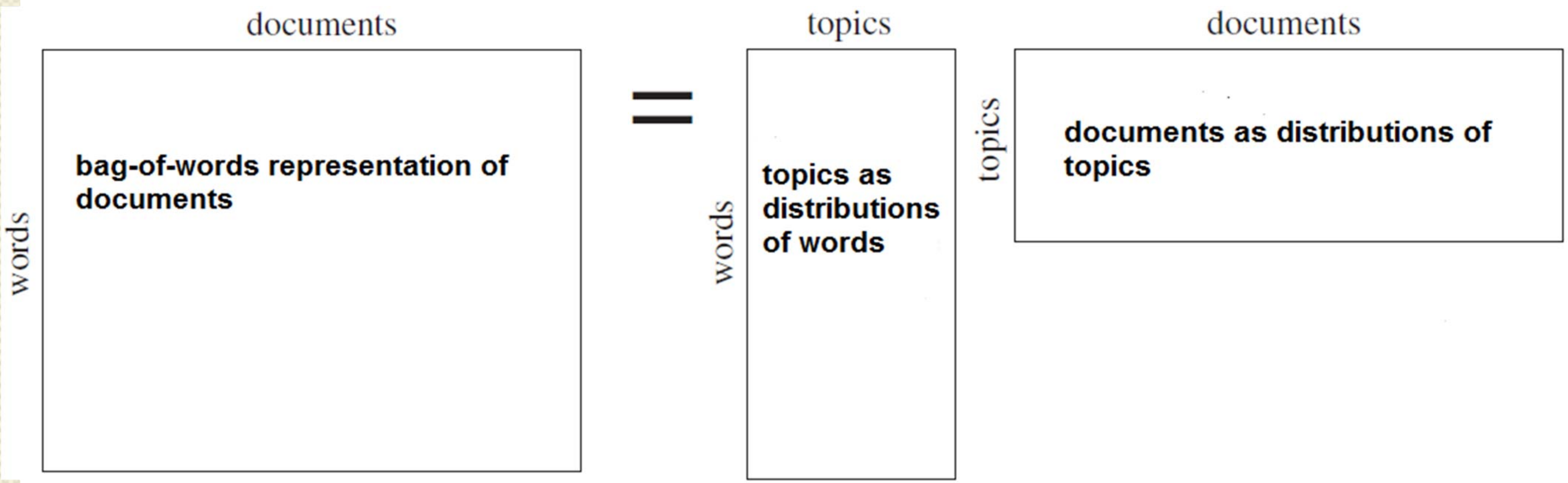
Topic models

- Topic models: a class of dimension reduction models on text (from words to topics)
- Bag-of-words representation of documents



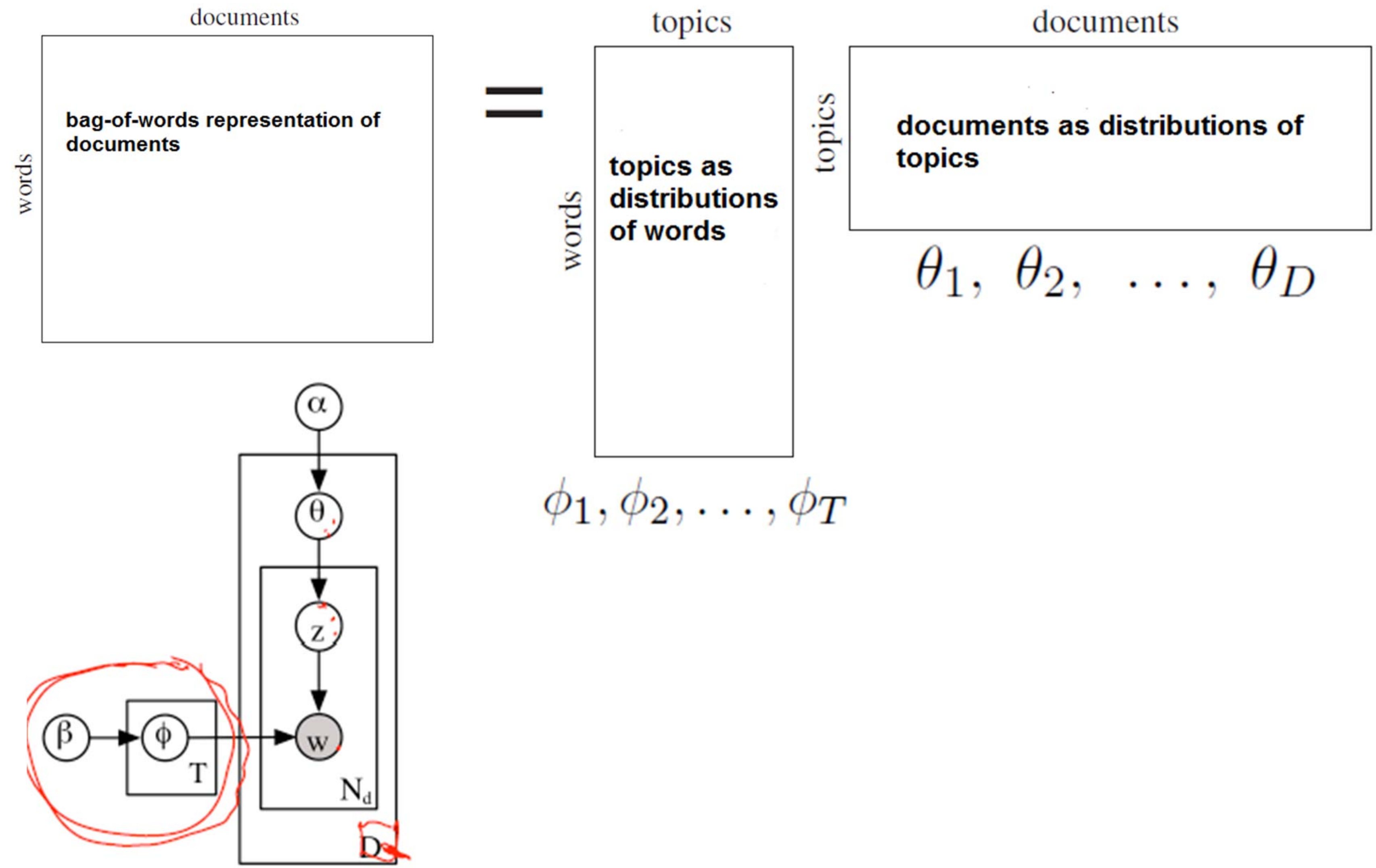
Topic models

- Topic models: a class of dimension reduction models on text (from words to topics)
- Bag-of-words representation of documents
- Topic models for representing documents

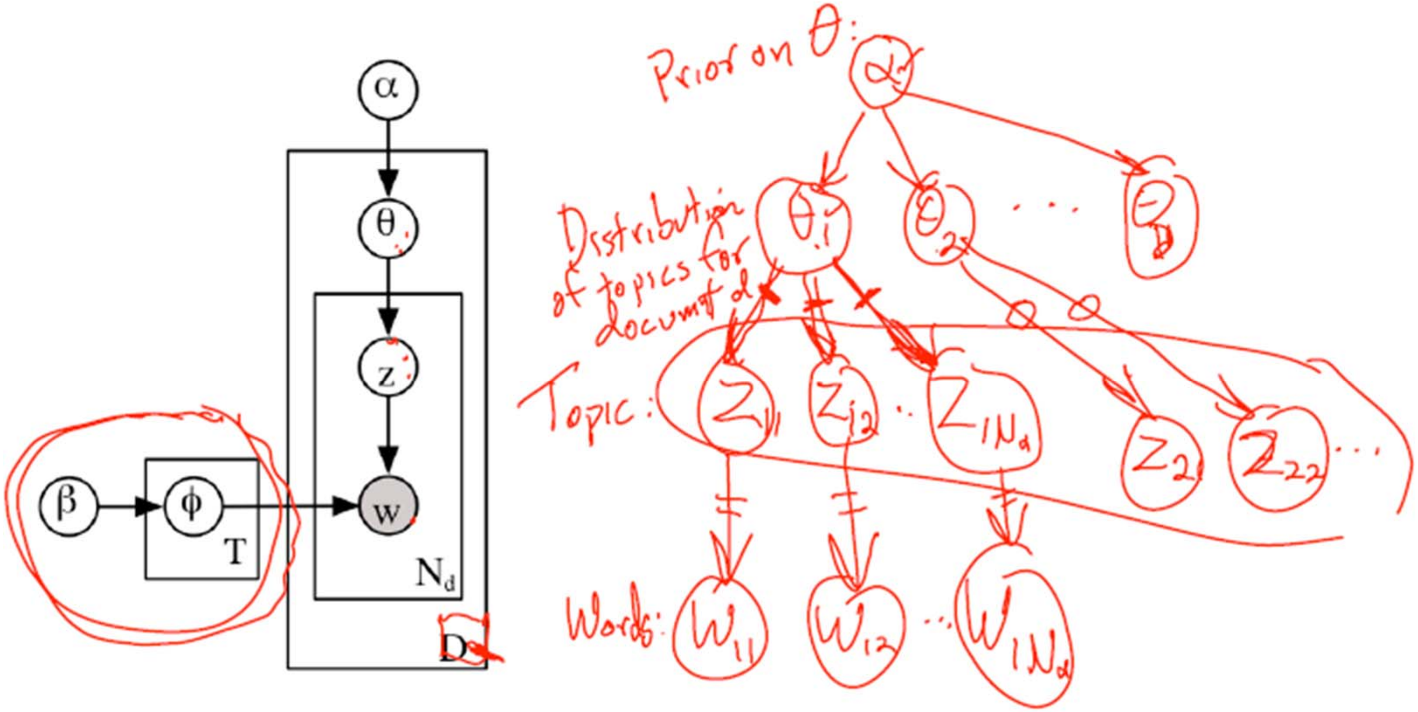


Latent Dirichlet allocation

- A fully Bayesian specification of topic models



Latent Dirichlet allocation



- Data: words on each documents
- Estimation: maximizing the data likelihood – difficult!

$$p(\mathbf{w}\,|\,\alpha,\beta) = \int p(\theta\,|\,\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n\,|\,\theta)\,p(w_n\,|\,z_n,\beta)\right)d\theta.$$