Understanding Topic Modeling: From Multivariate OLS to LDA

Salfo Bikienga
sbikienga@gmail.com

Columbus Machine Learners meetup

October 03, 2018

Agenda

- Introduction
- Prerequisite
- Non Negative Matrix Factorization
- Principal Component Analysis
- Latent Semantic Analysis
- Probabilistic Latent Semantic Analysis
- Latent Dirichlet Allocation
- Take home message

Introduction

Introduction

- Topic modeling methods are dimension reduction methods.

Introduction

- Topic modeling methods are dimension reduction methods.
- Generaly useful for:

Introduction

- Topic modeling methods are dimension reduction methods.
- Generaly useful for:
  - document clustering;

Introduction

- Topic modeling methods are dimension reduction methods.
- Generaly useful for:
  - document clustering;
  - document classification;

Introduction

- Topic modeling methods are dimension reduction methods.
- Generaly useful for:
  - document clustering;
  - document classification;
  - regression type of analysis;

Introduction

- Topic modeling methods are dimension reduction methods.
- Generaly useful for:
  - document clustering;
  - document classification;
  - regression type of analysis;
  - ...

Introduction

- The goal is to understand LDA through the lens of OLS.

Introduction

- The goal is to understand LDA through the lens of OLS.
- LDA is a Bayesian approach to pLSA.

Introduction

- The goal is to understand LDA through the lens of OLS.
- LDA is a Bayesian approach to pLSA.
- pLSA is a maximum likelihood approach to LSA.

Introduction

- The goal is to understand LDA through the lens of OLS.
- LDA is a Bayesian approach to pLSA.
- pLSA is a maximum likelihood approach to LSA.
- LSA is equivalent to PCA

Introduction

- The goal is to understand LDA through the lens of OLS.
- LDA is a Bayesian approach to pLSA.
- pLSA is a maximum likelihood approach to LSA.
- LSA is equivalent to PCA
- PCA is a matrix factorization algorithm (MF).

Introduction

- The goal is to understand LDA through the lens of OLS.
- LDA is a Bayesian approach to pLSA.
- pLSA is a maximum likelihood approach to LSA.
- LSA is equivalent to PCA
- PCA is a matrix factorization algorithm (MF).
- MF is an application of OLS.

Introduction

- The goal is to understand LDA through the lens of OLS.
- LDA is a Bayesian approach to pLSA.
- pLSA is a maximum likelihood approach to LSA.
- LSA is equivalent to PCA
- PCA is a matrix factorization algorithm (MF).
- MF is an application of OLS.
- The general idea of these algorithms is that:

$$W_{D\times V} \simeq Z_{D\times K} B_{K\times V}$$

where $K << V$

Introduction: practical example

Collapse a $W_{596\times1034}$ words counts into a $Z_{596\times2}$ matrix:

Table 1: Example of topics distributions when K = 2

	Topic.1	Topic.2
Alabama_2001_D_1.txt	0.75	0.25
Alabama_2002_D_2.txt	0.65	0.35
Alabama_2003_R_3.txt	0.26	0.74
Alabama_2004_R_4.txt	0.38	0.62
Alabama_2005_R_5.txt	0.50	0.50
Alabama_2006_R_6.txt	0.45	0.55

Z

Table 2: Words relative importance when K = 2

	Topic.1	Topic.2
abil	0.0004	0.001
abus	0.001	0.0004
academ	0.001	0.0000
acceler	0.0004	0.0000
accept	0.0002	0.001
access	0.003	0.0000
accomplish	0.001	0.001
accord	0.0000	0.001
account	0.001	0.002
achiev	0.003	0.001

B^T

Introduction: practical example

Table 3: List of words ordered by their relative importance for their respective topics. The list is used to infer the meaning of the topic.

Topic 1	Topic 2
school	budget
educ	fund
work	govern
help	peopl
econom	million
children	work
famili	make
health	public
busi	propos
nation	servic
make	chang
creat	program
student	know
teach	spend
invest	come

Prerequisite

Prerequisite: basic rules

- The Bayes rule:

$$p(B|Y) = \frac{p(Y|B) * p(B)}{p(Y)} \propto p(Y|B) * p(B)$$

Prerequisite: basic rules

- The Bayes rule:

$$p(B|Y) = \frac{p(Y|B) * p(B)}{p(Y)} \propto p(Y|B) * p(B)$$

- Matrix product rule:

$$\begin{pmatrix} 4 & 3 & 1 \\ 2 & 4 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 3 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix}$$

Prerequisite: basic rules

- The Bayes rule:

$$p(B|Y) = \frac{p(Y|B) * p(B)}{p(Y)} \propto p(Y|B) * p(B)$$

- Matrix product rule:

$$\begin{pmatrix} 4 & 3 & 1 \\ 2 & 4 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 3 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix}$$

- Transpose of a matrix product:

$$(AB)^T = B^T A^T$$

Prerequisite: Simple OLS

- Extended form:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \\ \vdots & \vdots \\ X_{n,1} & X_{n,2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Prerequisite: Simple OLS

- Extended form:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \\ \vdots & \vdots \\ X_{n,1} & X_{n,2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- Matrix form:

Prerequisite: Simple OLS

- Extended form:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \\ \vdots & \vdots \\ X_{n,1} & X_{n,2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- Matrix form:
  - $y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$

Prerequisite: Simple OLS

- Extended form:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \\ \vdots & \vdots \\ X_{n,1} & X_{n,2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- Matrix form:
  - $y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$
  - Assuming $X^T\epsilon = 0$,

$$X^T y = X^T X\beta$$

Prerequisite: Simple OLS

- Extended form:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \\ \vdots & \vdots \\ X_{n,1} & X_{n,2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- Matrix form:
  - $y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$
  - Assuming $X^T\epsilon = 0$,

$$X^T y = X^T X\beta$$

  - Assuming $X^T X$ invertible, $(X^T X)^{-1} X^T y = \hat{\beta}$

Prerequisite: Multivariate OLS

- Extended form:

$$\begin{pmatrix} y_{1,1} & y_{1,2} & y_{1,3} \\ y_{2,1} & y_{2,2} & y_{1,3} \\ \vdots & \vdots & \vdots \\ y_{n,1} & y_{n,2} & y_{1,3} \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \\ \vdots & \vdots \\ X_{n,1} & X_{n,2} \end{pmatrix} \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \epsilon_{1,3} \\ \epsilon_{2,1} & \epsilon_{2,2} & \epsilon_{2,3} \\ \vdots & & \\ \epsilon_{n,1} & \epsilon_{n,2} & \epsilon_{n,3} \end{pmatrix}$$

Prerequisite: Multivariate OLS

- Extended form:

$$\begin{pmatrix} y_{1,1} & y_{1,2} & y_{1,3} \\ y_{2,1} & y_{2,2} & y_{1,3} \\ \vdots & \vdots & \vdots \\ y_{n,1} & y_{n,2} & y_{1,3} \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \\ \vdots & \vdots \\ X_{n,1} & X_{n,2} \end{pmatrix} \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \epsilon_{1,3} \\ \epsilon_{2,1} & \epsilon_{2,2} & \epsilon_{2,3} \\ \vdots & & \\ \epsilon_{n,1} & \epsilon_{n,2} & \epsilon_{n,3} \end{pmatrix}$$

- Matrix form:

Prerequisite: Multivariate OLS

- Extended form:

$$\begin{pmatrix} y_{1,1} & y_{1,2} & y_{1,3} \\ y_{2,1} & y_{2,2} & y_{1,3} \\ \vdots & \vdots & \vdots \\ y_{n,1} & y_{n,2} & y_{1,3} \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \\ \vdots & \vdots \\ X_{n,1} & X_{n,2} \end{pmatrix} \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \epsilon_{1,3} \\ \epsilon_{2,1} & \epsilon_{2,2} & \epsilon_{2,3} \\ \vdots & & \\ \epsilon_{n,1} & \epsilon_{n,2} & \epsilon_{n,3} \end{pmatrix}$$

- Matrix form:
  - $Y_{n\times q} = X_{n\times p} B_{p\times q} + \epsilon_{n\times q}$

Prerequisite: Multivariate OLS

- Extended form:

$$\begin{pmatrix} y_{1,1} & y_{1,2} & y_{1,3} \\ y_{2,1} & y_{2,2} & y_{1,3} \\ \vdots & \vdots & \vdots \\ y_{n,1} & y_{n,2} & y_{1,3} \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \\ \vdots & \vdots \\ X_{n,1} & X_{n,2} \end{pmatrix} \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \epsilon_{1,3} \\ \epsilon_{2,1} & \epsilon_{2,2} & \epsilon_{2,3} \\ \vdots & & \\ \epsilon_{n,1} & \epsilon_{n,2} & \epsilon_{n,3} \end{pmatrix}$$

- Matrix form:
  - $Y_{n\times q} = X_{n\times p}B_{p\times q} + \epsilon_{n\times q}$
  - And

$$\hat{B} = (X^TX)^{-1}X^TY$$

Prerequisite: Multivariate OLS

- Extended form:

$$\begin{pmatrix} y_{1,1} & y_{1,2} & y_{1,3} \\ y_{2,1} & y_{2,2} & y_{1,3} \\ \vdots & \vdots & \vdots \\ y_{n,1} & y_{n,2} & y_{1,3} \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ \vdots & \vdots \\ x_{n,1} & x_{n,2} \end{pmatrix} \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \epsilon_{1,3} \\ \epsilon_{2,1} & \epsilon_{2,2} & \epsilon_{2,3} \\ \vdots & & \\ \epsilon_{n,1} & \epsilon_{n,2} & \epsilon_{n,3} \end{pmatrix}$$

- Matrix form:
  - $Y_{n\times q} = X_{n\times p} B_{p\times q} + \epsilon_{n\times q}$
  - And
$$\hat{B} = (X^T X)^{-1} X^T Y$$

- Note: no distributional assumption is required.

Prerequisite: Maximum Likelihood Estimation

- Assume $Y_{i_{q\times 1}} \overset{iid}{\sim} Normal(B_{q\times p}X_{i_{p\times 1}}, \alpha^{-1}I_q)$

Prerequisite: Maximum Likelihood Estimation

- Assume $Y_{i_{q\times 1}} \overset{iid}{\sim} Normal(B_{q\times p}X_{i_{p\times 1}}, \alpha^{-1}I_q)$
- $P(Y_i|B,\alpha) = Cexp\{-\frac{\alpha}{2}(Y_i - B^TX_i)^T(Y_i - B^TX_i)\}$

Prerequisite: Maximum Likelihood Estimation

- Assume $Y_{i_{q\times 1}} \overset{iid}{\sim} Normal(B_{q\times p}X_{i_{p\times 1}}, \alpha^{-1}I_q)$
- $P(Y_i|B,\alpha) = Cexp\{-\frac{\alpha}{2}(Y_i - B^TX_i)^T(Y_i - B^TX_i)\}$
- $L(B,\alpha) = C^Nexp\{-\frac{\alpha}{2}\sum_{i=1}^{n}(Y_i - B^TX_i)^T(Y_i - B^TX_i)\}$

Prerequisite: Maximum Likelihood Estimation

- Assume $Y_{i_{q\times 1}} \overset{iid}{\sim} Normal(B_{q\times p}X_{i_{p\times 1}}, \alpha^{-1}I_q)$
- $P(Y_i|B,\alpha) = Cexp\{-\frac{\alpha}{2}(Y_i - B^TX_i)^T(Y_i - B^TX_i)\}$
- $L(B,\alpha) = C^Nexp\{-\frac{\alpha}{2}\sum_{i=1}^{n}(Y_i - B^TX_i)^T(Y_i - B^TX_i)\}$
- $\ell(B,\alpha) = Nlog(C) - \frac{\alpha}{2}\sum_{i=1}^{n}(Y_i - B^TX_i)^T(Y_i - B^TX_i)$

Prerequisite: Maximum Likelihood Estimation

- Assume $Y_{i_{q\times 1}} \overset{iid}{\sim} Normal(B_{q\times p}X_{i_{p\times 1}}, \alpha^{-1}I_q)$
- $P(Y_i|B,\alpha) = Cexp\{-\frac{\alpha}{2}(Y_i - B^TX_i)^T(Y_i - B^TX_i)\}$
- $L(B,\alpha) = C^N exp\{-\frac{\alpha}{2}\sum_{i=1}^{n}(Y_i - B^TX_i)^T(Y_i - B^TX_i)\}$
- $\ell(B,\alpha) = Nlog(C) - \frac{\alpha}{2}\sum_{i=1}^{n}(Y_i - B^TX_i)^T(Y_i - B^TX_i)$
- In matrix form,

Prerequisite: Maximum Likelihood Estimation

- Assume $Y_{i_{q\times 1}} \overset{iid}{\sim} Normal(B_{q\times p}X_{i_{p\times 1}}, \alpha^{-1}I_q)$
- $P(Y_i|B,\alpha) = Cexp\{-\frac{\alpha}{2}(Y_i - B^TX_i)^T(Y_i - B^TX_i)\}$
- $L(B,\alpha) = C^N exp\{-\frac{\alpha}{2}\sum_{i=1}^{n}(Y_i - B^TX_i)^T(Y_i - B^TX_i)\}$
- $\ell(B,\alpha) = Nlog(C) - \frac{\alpha}{2}\sum_{i=1}^{n}(Y_i - B^TX_i)^T(Y_i - B^TX_i)$
- In matrix form,
  - $\ell(B,\alpha) = Nlog(C) - \frac{\alpha}{2}(Y - XB)^T(Y - XB)$

Prerequisite: Maximum Likelihood Estimation

- Assume $Y_{i_{q\times 1}} \overset{iid}{\sim} Normal(B_{q\times p}X_{i_{p\times 1}}, \alpha^{-1}I_q)$
- $P(Y_i|B,\alpha) = Cexp\{-\frac{\alpha}{2}(Y_i - B^TX_i)^T(Y_i - B^TX_i)\}$
- $L(B,\alpha) = C^N exp\{-\frac{\alpha}{2}\sum_{i=1}^{n}(Y_i - B^TX_i)^T(Y_i - B^TX_i)\}$
- $\ell(B,\alpha) = Nlog(C) - \frac{\alpha}{2}\sum_{i=1}^{n}(Y_i - B^TX_i)^T(Y_i - B^TX_i)$
- In matrix form,
  - $\ell(B,\alpha) = Nlog(C) - \frac{\alpha}{2}(Y - XB)^T(Y - XB)$
  - $\ell(B,\alpha) \simeq -\frac{\alpha}{2}[B^T(X^TX)B - B^T(X^TY)]$

Prerequisite: Maximum Likelihood Estimation

- Assume $Y_{i_{q\times 1}} \overset{iid}{\sim} Normal(B_{q\times p}X_{i_{p\times 1}}, \alpha^{-1}I_q)$
- $P(Y_i|B,\alpha) = Cexp\{-\frac{\alpha}{2}(Y_i - B^T X_i)^T(Y_i - B^T X_i)\}$
- $L(B,\alpha) = C^N exp\{-\frac{\alpha}{2}\sum_{i=1}^{n}(Y_i - B^T X_i)^T(Y_i - B^T X_i)\}$
- $\ell(B,\alpha) = Nlog(C) - \frac{\alpha}{2}\sum_{i=1}^{n}(Y_i - B^T X_i)^T(Y_i - B^T X_i)$
- In matrix form,
  - $\ell(B,\alpha) = Nlog(C) - \frac{\alpha}{2}(Y - XB)^T(Y - XB)$
  - $\ell(B,\alpha) \simeq -\frac{\alpha}{2}[B^T(X^T X)B - B^T(X^T Y)]$
  - $$\frac{\partial \ell}{\partial B} = 0 \Longrightarrow (X^T X)B - (X^T Y) = 0$$

Prerequisite: Multivariate Bayesian Regression

- Assume B has a gaussian prior, i.e. $B \sim Normal(m_0, V_0)$

Prerequisite: Multivariate Bayesian Regression

- Assume B has a gaussian prior, i.e. $B \sim Normal(m_0, V_0)$
- By bayes rule:

$$P(B|Y) \propto P(Y|B, \alpha)P(B|m_0, V_0)$$

Prerequisite: Multivariate Bayesian Regression

- Assume B has a gaussian prior, i.e. $B \sim Normal(m_0, V_0)$
- By bayes rule:

$$P(B|Y) \propto P(Y|B, \alpha)P(B|m_0, V_0)$$

- It can be shown that:

$$B|Y \sim Normal(m, V)$$

Prerequisite: Multivariate Bayesian Regression

- Assume B has a gaussian prior, i.e. $B \sim Normal(m_0, V_0)$
- By bayes rule:

$$P(B|Y) \propto P(Y|B, \alpha)P(B|m_0, V_0)$$

- It can be shown that:

$$B|Y \sim Normal(m, V)$$

  - where $m = (X^T X + V_0)^{-1}(X^T Y + V_0 m_0)$

Prerequisite: Multivariate Bayesian Regression

- Assume B has a gaussian prior, i.e. $B \sim Normal(m_0, V_0)$
- By bayes rule:

$$P(B|Y) \propto P(Y|B, \alpha)P(B|m_0, V_0)$$

- It can be shown that:

$$B|Y \sim Normal(m, V)$$

  - where $m = (X^T X + V_0)^{-1}(X^T Y + V_0 m_0)$
  - i.e. $m = (X^T X + V_0)^{-1}[(X^T X)(X^T X)^{-1} X^T Y + V_0 m_0]$

Prerequisite: Multivariate Bayesian Regression

- Assume B has a gaussian prior, i.e. $B \sim Normal(m_0, V_0)$
- By bayes rule:

$$P(B|Y) \propto P(Y|B, \alpha)P(B|m_0, V_0)$$

- It can be shown that:

$$B|Y \sim Normal(m, V)$$

  - where $m = (X^T X + V_0)^{-1}(X^T Y + V_0 m_0)$
  - i.e. $m = (X^T X + V_0)^{-1}[(X^T X)(X^T X)^{-1} X^T Y + V_0 m_0]$
  - and
$$m = (X^T X + V_0)^{-1}[(X^T X)\hat{B} + V_0 m_0]$$

Prerequisite: Bayesian Regression vs OLS

Prerequesite take home,

Method	Parameter estimate
OLS	$\hat{B}_{ols} = (X^T X)^{-1} X^T Y$
MLE	$\hat{B}_{mle} = (X^T X)^{-1} X^T Y$
Bayesian	$\hat{B}_{bayes} = (X^T X + V_0)^{-1}[(X^T X)\hat{B}_{ols} + V_0 m_0]$

Prerequisite: Bayesian Regression vs OLS

Prerequesite take home,

Method	Parameter estimate	Topic Models
OLS	$\hat{B}_{ols} = (X^T X)^{-1} X^T Y$	MF/NMF
MLE	$\hat{B}_{mle} = (X^T X)^{-1} X^T Y$	pLSA
Bayesian	$\hat{B}_{bayes} = (X^T X + V_0)^{-1}[(X^T X)\hat{B}_{ols} + V_0 m_0]$	LDA

Non Negative Matrix Factorization (NMF)

NMF: Matrix Factorization (MF)

- There are several MF algorithms, mostly used for two purposes:

NMF: Matrix Factorization (MF)

- There are several MF algorithms, mostly used for two purposes:
  - To solve linear systems (e.g.: LU, QR decompositions);

NMF: Matrix Factorization (MF)

- There are several MF algorithms, mostly used for two purposes:
  - To solve linear systems (e.g.: LU, QR decompositions);
  - For statistical analysis (e.g.: Factor Analysis, PCA/LSA).

NMF: Matrix Factorization (MF)

- There are several MF algorithms, mostly used for two purposes:
  - To solve linear systems (e.g.: LU, QR decompositions);
  - For statistical analysis (e.g.: Factor Analysis, PCA/LSA).
- The general idea of MF is:

NMF: Matrix Factorization (MF)

- There are several MF algorithms, mostly used for two purposes:
  - To solve linear systems (e.g.: LU, QR decompositions);
  - For statistical analysis (e.g.: Factor Analysis, PCA/LSA).
- The general idea of MF is:
  - Let $W_{D \times V}$ be a matrix of dimension $D \times V$;

NMF: Matrix Factorization (MF)

- There are several MF algorithms, mostly used for two purposes:
  - To solve linear systems (e.g.: LU, QR decompositions);
  - For statistical analysis (e.g.: Factor Analysis, PCA/LSA).
- The general idea of MF is:
  - Let $W_{D \times V}$ be a matrix of dimension $D \times V$;
  - then

$$W_{D \times V} \simeq Z_{D \times K} B_{K \times V}$$

NMF: Matrix Factorization (MF)

- There are several MF algorithms, mostly used for two purposes:
  - To solve linear systems (e.g.: LU, QR decompositions);
  - For statistical analysis (e.g.: Factor Analysis, PCA/LSA).
- The general idea of MF is:
  - Let $W_{D\times V}$ be a matrix of dimension $D \times V$;
  - then

$$W_{D\times V} \simeq Z_{D\times K} B_{K\times V}$$

  - K is an arbitrary number.

NMF: Matrix Factorization (MF)

Table 6: Example matrix of words counts

	college	education	family	health	medicaid
document.1	4	6	0	2	2
document.2	0	0	4	8	12
document.3	6	9	1	5	6
document.4	2	3	3	7	10
document.5	0	0	3	6	9
document.6	4	6	1	4	5

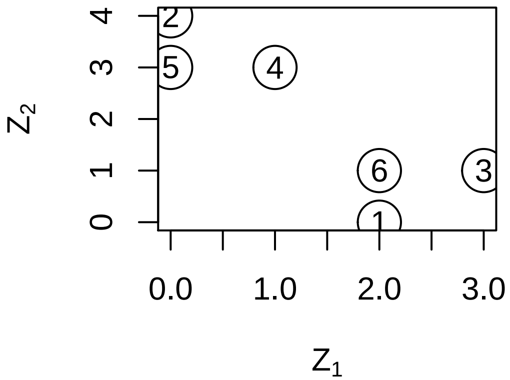
NMF: Matrix Factorization (MF)

Example:

$$\underbrace{\begin{bmatrix} 4 & 6 & 0 & 2 & 2 \\ 0 & 0 & 4 & 8 & 12 \\ 6 & 9 & 1 & 5 & 6 \\ 2 & 3 & 3 & 7 & 10 \\ 0 & 0 & 3 & 6 & 9 \\ 4 & 6 & 1 & 4 & 5 \end{bmatrix}}_{\mathbf{W_{6\times 5}}} \simeq \underbrace{\begin{bmatrix} 2 & 0 \\ 0 & 4 \\ 3 & 1 \\ 1 & 3 \\ 0 & 3 \\ 2 & 1 \end{bmatrix}}_{\mathbf{Z_{6\times 2}}} \underbrace{\begin{bmatrix} 2 & 3 & 0 & 1 & 1 \\ 0 & 0 & 1 & 2 & 3 \end{bmatrix}}_{\mathbf{B_{2\times 5}}}$$

NMF: Matrix Factorization (MF)

Scatterplot based on the Z matr



NMF: Matrix Factorization (MF)

MF: Iterative Multivariate Least Square algorithm

- Write:

$$W_{D \times V} = Z_{D \times K} B_{K \times V} + \epsilon_{D \times V}$$

NMF: Matrix Factorization (MF)

MF: Iterative Multivariate Least Square algorithm

- Write:

$$W_{D\times V} = Z_{D\times K} B_{K\times V} + \epsilon_{D\times V}$$

- From multivariate regression, we know:

$$\hat{B}_{K\times V} = (Z^T Z)^{-1} Z^T W$$

NMF: Matrix Factorization (MF)

MF: Iterative Multivariate Least Square algorithm

- Write:

$$W_{D\times V} = Z_{D\times K} B_{K\times V} + \epsilon_{D\times V}$$

- From multivariate regression, we know:

$$\hat{B}_{K\times V} = (Z^T Z)^{-1} Z^T W$$

- But, we do not have Z; however, we can write:

$$\hat{Z}_{D\times K} = W B^T [B B^T]^{-1}$$

NMF: Matrix Factorization (MF)

MF: Iterative Multivariate Least Square algorithm

- Write:

$$W_{D\times V} = Z_{D\times K} B_{K\times V} + \epsilon_{D\times V}$$

- From multivariate regression, we know:

$$\hat{B}_{K\times V} = (Z^T Z)^{-1} Z^T W$$

- But, we do not have Z; however, we can write:

$$\hat{Z}_{D\times K} = W B^T [B B^T]^{-1}$$

- Initialize random Z, and itteratively solve for B and Z.

Non Negative Matrix Factorization (NMF)

Non Negative Matrix Factorization

- Impose constraints such that: $Z_{i,j} \geq 0$, and $B_{i,j} \geq 0$

$$W_{D \times V} \simeq Z^{nmf} B^{nmf}$$

Non Negative Matrix Factorization (NMF)

Non Negative Matrix Factorization

- Impose constraints such that: $Z_{i,j} \geq 0$, and $B_{i,j} \geq 0$

$$W_{D \times V} \simeq Z^{nmf} B^{nmf}$$

- Let $D_{W_{d,d}} = \sum_{v=1}^{V} W_{d,v}$ and $D_{B_{k,k}} = \sum_{v=1}^{V} B_{k,v}$ be some normalizing matrices.

Non Negative Matrix Factorization (NMF)

Non Negative Matrix Factorization

- Impose constraints such that: $Z_{i,j} \geq 0$, and $B_{i,j} \geq 0$

$$W_{D \times V} \simeq Z^{nmf} B^{nmf}$$

- Let $D_{W_{d,d}} = \sum_{v=1}^{V} W_{d,v}$ and $D_{B_{k,k}} = \sum_{v=1}^{V} B_{k,v}$ be some normalizing matrices.
- Then, Z^* and B^* can be interpreted as probabilities:

$$D_W^{-1} W = \left[D_W^{-1} Z D_B \right] \left[D_B^{-1} B \right]$$

$$\Longleftrightarrow$$

$$W^* = Z^* B^*$$

NMF: Matrix Factorization (MF)

Observation:

- $$\hat{B}_{K \times V} = \left[Z^T Z \right]^{-1} Z^T W = P_{K \times D} W_{D \times V}$$
- $$\hat{B} = \begin{pmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,V} \\ B_{2,1} & B_{2,2} & \cdots & B_{1,V} \\ \vdots & \vdots & \ddots & \vdots \\ B_{K,1} & B_{K,2} & \cdots & B_{K,V} \end{pmatrix}$$
- $$\hat{B}_{k,v} = \sum_{d=1}^{D} P_{k,d} W_{d,v}$$

NMF: Matrix Factorization (MF)

Observation:

- $$\hat{Z}_{D\times K} = WB^T \left[BB^T\right]^{-1} = W_{D\times V} Q_{V\times K}$$
- $$\hat{Z} = \begin{pmatrix} Z_{1,1} & Z_{1,2} & \cdots & Z_{1,K} \\ Z_{2,1} & Z_{2,2} & \cdots & Z_{1,K} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{D,1} & Z_{D,2} & \cdots & {}_{D,K} \end{pmatrix}$$
- $$\hat{Z}_{d,k} = \sum_{v=1}^{V} Q_{v,k} W_{d,v}$$

Principal Component Analysis (PCA)

PCA: Spectral decomposition

- PCA is MF with two additional constraints:

PCA: Spectral decomposition

- PCA is MF with two additional constraints:
  - We want Z to be non correlated (orthogonal($\perp$));

PCA: Spectral decomposition

- PCA is MF with two additional constraints:
  - We want Z to be non correlated (orthogonal($\perp$));
  - We also want to preserve the variance of the W matrix.

PCA: Spectral decomposition

- PCA is MF with two additional constraints:
  - We want Z to be non correlated (orthogonal($\perp$));
  - We also want to preserve the variance of the W matrix.
- Solution: find an $\perp$ matrix $\tilde{B}$ such that $Z = W\tilde{B}$ is $\perp$.

PCA: Spectral decomposition

- PCA is MF with two additional constraints:
  - We want Z to be non correlated (orthogonal($\perp$));
  - We also want to preserve the variance of the W matrix.
- Solution: find an $\perp$ matrix $\tilde{B}$ such that $Z = W\tilde{B}$ is $\perp$.
- Observe that if $Z = W\tilde{B}$, then:

$$
\begin{aligned}
C_Z &= \frac{1}{n-1} Z^T Z \\
&= \frac{1}{n-1} \left[\tilde{B}^T W^T W \tilde{B} \right] \\
&= \tilde{B}^T \left[\frac{1}{n-1} W^T W \right] \tilde{B} \\
&= \tilde{B}^T C_W \tilde{B}
\end{aligned}
$$

PCA: Spectral decomposition

- Thus:

$$C_Z = \tilde{B}^T C_W \tilde{B}$$

PCA: Spectral decomposition

- Thus:
$$C_Z = \tilde{B}^T C_W \tilde{B}$$
- **Theorem:** If A is symmetric, there is an orthonormal matrix E such that $A = EDE^T$, where D is a diagonal matrix.

PCA: Spectral decomposition

- Thus:
$$C_Z = \tilde{B}^T C_W \tilde{B}$$
- **Theorem:** If A is symmetric, there is an orthonormal matrix E such that $A = EDE^T$, where D is a diagonal matrix.
- This theorem (Spectral decomposition) is all we need for PCA.

PCA: Spectral decomposition

- Thus:
$$C_Z = \tilde{B}^T C_W \tilde{B}$$
- **Theorem:** If A is symmetric, there is an orthonormal matrix E such that $A = EDE^T$, where D is a diagonal matrix.
- This theorem (Spectral decomposition) is all we need for PCA.
- Translation:

PCA: Spectral decomposition

- Thus:
$$C_Z = \tilde{B}^T C_W \tilde{B}$$
- **Theorem:** If A is symmetric, there is an orthonormal matrix E such that $A = EDE^T$, where D is a diagonal matrix.
- This theorem (Spectral decomposition) is all we need for PCA.
- Translation:
  - Compute the C_W from the data matrix (W);

PCA: Spectral decomposition

- Thus:
$$C_Z = \tilde{B}^T C_W \tilde{B}$$
- **Theorem:** If A is symmetric, there is an orthonormal matrix E such that $A = EDE^T$, where D is a diagonal matrix.
- This theorem (Spectral decomposition) is all we need for PCA.
- Translation:
  - Compute the C_W from the data matrix (W);
  - Use eigen-decomposition to get E, and use E as $\tilde{B}$;

PCA: Spectral decomposition

- Thus:
$$C_Z = \tilde{B}^T C_W \tilde{B}$$
- **Theorem:** If A is symmetric, there is an orthonormal matrix E such that $A = EDE^T$, where D is a diagonal matrix.
- This theorem (Spectral decomposition) is all we need for PCA.
- Translation:
  - Compute the C_W from the data matrix (W);
  - Use eigen-decomposition to get E, and use E as $\tilde{B}$;
  - Then compute $Z = WE = W\tilde{B}$

PCA: Spectral decomposition

- To check if Z is $\perp$, use the theorem and set $\tilde{B} = E$,

$$\begin{aligned}
C_Z &= \tilde{B}^T C_W \tilde{B} \\
&= E^T \left[EDE^T \right] E \\
&= E^T EDE^T E \\
&= D
\end{aligned}$$

PCA: Spectral decomposition

- To check if Z is $\perp$, use the theorem and set $\tilde{B} = E$,

$$\begin{aligned} C_Z &= \tilde{B}^T C_W \tilde{B} \\ &= E^T \left[EDE^T \right] E \\ &= E^T EDE^T E \\ &= D \end{aligned}$$

- **Definition:** The total variance is the trace of the covariance matrix

$$\begin{aligned} tr(C_Z) &= tr(D) \\ &= tr(\tilde{B}^T C_W \tilde{B}) \\ &= tr(E^T C_W E) \\ &= tr(EE^T C_W) \\ &= tr(C_W) \end{aligned}$$

PCA: Spectral decomposition

- As a dimension reduction method, we hope that there is a $K << V$ such that $\sum_{k=1}^{K} d_{k,k} \simeq tr(C_W)$; in which case, $Z_{D\times K} \simeq W_{D\times V} E_{V\times K}$ approximates $W_{D\times V}$.

PCA: Spectral decomposition

- As a dimension reduction method, we hope that there is a $K << V$ such that $\sum_{k=1}^{K} d_{k,k} \simeq tr(C_W)$; in which case, $Z_{D\times K} \simeq W_{D\times V} E_{V\times K}$ approximates $W_{D\times V}$.
- Then, we can approximately retrieve $W_{D\times V}$ by writing:

$$
\begin{aligned}
Z_{D\times K} E^T_{K\times V} &\simeq W_{D\times V} E_{V\times K} E^T_{K\times V} \\
&\iff \\
W_{D\times V} &\simeq Z_{D\times K} E^T_{K\times V} \\
&= ZB
\end{aligned}
$$

PCA: Spectral decomposition

- As a dimension reduction method, we hope that there is a $K << V$ such that $\sum_{k=1}^{K} d_{k,k} \simeq tr(C_W)$; in which case, $Z_{D\times K} \simeq W_{D\times V} E_{V\times K}$ approximates $W_{D\times V}$.
- Then, we can approximately retrieve $W_{D\times V}$ by writing:

$$\begin{aligned} Z_{D\times K} E^T_{K\times V} &\simeq W_{D\times V} E_{V\times K} E^T_{K\times V} \\ &\iff \\ W_{D\times V} &\simeq Z_{D\times K} E^T_{K\times V} \\ &= ZB \end{aligned}$$

- Where $B_{K\times V} = E^T_{K\times V}$ and $Z_{D\times K} = W_{D\times V} E_{V\times K}$

PCA: Singular Value Decomposition (SVD)

- SVD is a more general PCA algorithm.

PCA: Singular Value Decomposition (SVD)

- SVD is a more general PCA algorithm.
- SVD states that any matrix W can be decomposed as follows:

$$W_{D \times V} = U_{D \times D} S_{D \times V} V^T_{V \times V}$$

PCA: Singular Value Decomposition (SVD)

- SVD is a more general PCA algorithm.
- SVD states that any matrix W can be decomposed as follows:

$$W_{D \times V} = U_{D \times D} S_{D \times V} V^T_{V \times V}$$

- U, V are orthonormal matrices, i.e. $U^T U = UU^T = I_D$, $V^T V = VV^T = I_V$. S is a diagonal matrix containing the $r = min(D, V)$ singular values $\sigma_k \geq 0$ on the main diagonal, with 0s filling the rest of the matrix.

PCA: Singular Value Decomposition (SVD)

- By SVD, i.e.

$$W_{D \times V} = U_{D \times D} S_{D \times V} V^T_{V \times V}$$

PCA: Singular Value Decomposition (SVD)

- By SVD, i.e.

$$W_{D\times V} = U_{D\times D} S_{D\times V} V^T_{V\times V}$$

- If $W_{D\times V}$ are zero means V variables, the covariance matrix:

$$\begin{aligned} C_W &= \frac{1}{n-1} W^T W \\ &= \frac{1}{n-1} VSU^T USV^T \\ &= \frac{1}{n-1} VS^2 V^T \\ &= VDV^T \end{aligned}$$

PCA: Singular Value Decomposition (SVD)

- If there is a K such that $\sigma_{K+i} \simeq 0$, for $i = 1, 2, \cdots, V - K$, we can approximate $W_{D\times V}$, by

$$W_{D\times V} \simeq U_{D\times K} S_{K\times K} V^T_{K\times V}$$

PCA: Singular Value Decomposition (SVD)

- If there is a K such that $\sigma_{K+i} \simeq 0$, for $i = 1, 2, \cdots, V - K$, we can approximate $W_{D\times V}$, by

$$W_{D\times V} \simeq U_{D\times K} S_{K\times K} V^T_{K\times V}$$

- Along the spirit of $W \simeq ZB$, let's define $Z = US$, and $B = V^T$. Then, we can write:

$$W \simeq ZB$$

Latent Semantic Analysis (LSA)

LSA

- LSA is an application of SVD to a matrix of words counts.

LSA

- LSA is an application of SVD to a matrix of words counts.
- As such, LSA is exactly another application of PCA.

LSA

- LSA is an application of SVD to a matrix of words counts.
- As such, LSA is exactly another application of PCA.
- Example

Probabilistic Latent Semantic Analysis (pLSA)

pLSA

- For statisticians, LSA has two major problems:

pLSA

- For statisticians, LSA has two major problems:
  - It does not account for the fact that text data are count data.

pLSA

- For statisticians, LSA has two major problems:
  - It does not account for the fact that text data are count data.
  - It does not assume any distribution for the data.

pLSA

- For statisticians, LSA has two major problems:
  - It does not account for the fact that text data are count data.
  - It does not assume any distribution for the data.
- pLSA was proposed to address these concerns

Probabilistic Latent Semantic Analysis (pLSA)

- Assume $p(w_v|d_i)$ is the probability of observing the word w_v in the document d_i.

Probabilistic Latent Semantic Analysis (pLSA)

- Assume $p(w_v|d_i)$ is the probability of observing the word w_v in the document d_i.
- Then:

$$\begin{aligned} p(w_v|d_i) &= \sum_{z\in Z} p(w_v, z|d_i) \\ &= \sum_{z\in Z} p(w_v|z, d_i)p(z|d_i) \\ &= \sum_{z\in Z} p(w_v|z)p(z|d_i) \end{aligned}$$

Probabilistic Latent Semantic Analysis (pLSA)

- A document is a collection of $N_{d_i} = \sum_v^V n_{d_i, w_v}$ words, assumed independent. Therefore:

$$p(w_1, w_2, \cdots, w_V | d_i) = \prod_{v=1}^{V} p(w_v | d_i)^{n(d_i, w_v)}$$

Probabilistic Latent Semantic Analysis (pLSA)

- A document is a collection of $N_{d_i} = \sum_v^V n_{d_i,w_v}$ words, assumed independent. Therefore:

$$p(w_1, w_2, \cdots, w_V | d_i) = \prod_{v=1}^{V} p(w_v | d_i)^{n(d_i, w_v)}$$

- Assuming D independent documents,

$$L(\theta | W) = p(W | D) = \prod_{d=1}^{D} \prod_{v=1}^{V} p(w_v | d_i)^{n(d_i, w_v)}$$

$$\mathcal{L}(\theta | W) = \sum_{d=1}^{D} \sum_{v=1}^{V} n(d_i, w_v) log \left(\sum_{z \in Z} p(w_v | z) p(z | d_i) \right)$$

Probabilistic Latent Semantic Analysis (pLSA)

- $$p(z_k|d_i, w_v) = \frac{p(w_v|z_k)p(z_k|d_i)}{\sum_{l=1}^{K} p(w_v|z_l)p(z_l|d_i)}$$

Probabilistic Latent Semantic Analysis (pLSA)

- $$p(z_k|d_i, w_v) = \frac{p(w_v|z_k)p(z_k|d_i)}{\sum_{l=1}^{K} p(w_v|z_l)p(z_l|d_i)}$$
- $$p(w_v|z_k) = \frac{\sum_{d=1}^{D} n(d_i, w_v)p(z_k|d_i, w_v)}{\sum_{v=1}^{V}\sum_{d=1}^{D} n(d_i, w_v)p(z_k|d_i, w_v)}$$

Probabilistic Latent Semantic Analysis (pLSA)

- $$p(z_k|d_i, w_v) = \frac{p(w_v|z_k)p(z_k|d_i)}{\sum_{l=1}^{K} p(w_v|z_l)p(z_l|d_i)}$$
- $$p(w_v|z_k) = \frac{\sum_{d=1}^{D} n(d_i, w_v)p(z_k|d_i, w_v)}{\sum_{v=1}^{V} \sum_{d=1}^{D} n(d_i, w_v)p(z_k|d_i, w_v)}$$
- $$p(z_k|d_i) = \frac{\sum_{v=1}^{V} n(d_i, w_v)p(z_k|d_i, w_v)}{\sum_{k=1}^{K} \sum_{v=1}^{V} n(d_i, w_v)p(z_k|d_i, w_v)}$$

Probabilistic Latent Semantic Analysis (pLSA)

- $$p(w_v, d_i) = \sum_{Z} p(z)p(w_v|z)p(d_i|z) = \sum_{z_k=1}^{K} p(d_i|z_k)p(z_k)p(w_v|z_k)$$
- Let's define $U = [p(d_i|z_k)]_{D\times K}$, $V^T = [p(w_v|z_k)]_{K\times V}$, and $S = [p(z_k)]_{K\times K}$.
- Then, it follows that:

$$\begin{aligned}
[p(w_v, d_i)]_{D\times V} &= \sum_{z_k=1}^{K} p(d_i|z_k)p(z_k)p(w_v|z_k) \\
&= [p(d_i|z_k)]_{D\times K} [p(z_k)]_{K\times K} [p(w_v|z_k)]_{K\times V} \\
&= USV^T
\end{aligned}$$

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA)

- LDA is a Bayesian treatment of pLSA

$$p(z_k|d) = \theta_{d,k}$$
$$p(w_v|z_k) = \phi_{k,v}$$

$$\theta_d \sim Dirichlet_K(\alpha)$$
$$\phi_k \sim Dirichlet_V(\beta_k)$$

Latent Dirichlet Allocation (LDA)

- MCMC or Variational Bayes (VB) methods are used to approximate the posterior distribution for θ and ϕ.

Latent Dirichlet Allocation (LDA)

- MCMC or Variational Bayes (VB) methods are used to approximate the posterior distribution for θ and ϕ.
- By VB,

$$\theta_d | w_d, \tilde{\boldsymbol{\alpha}} \sim Dirichlet_K(\tilde{\boldsymbol{\alpha}}_{\boldsymbol{d}})$$

$$\phi_k | w, \tilde{\boldsymbol{\beta}} \sim Dirichlet_V(\tilde{\boldsymbol{\beta}}_{\boldsymbol{k}})$$

Latent Dirichlet Allocation (LDA)

$$E(z_{d,v,\cdot}) = exp(E(log(\theta_{d,\cdot})) + E(log(\phi_{\cdot,v})))$$

$$E(\theta_d|\tilde{\alpha_d}) = \frac{\alpha + \sum_{v=1}^{V} n_{d,v} \times E(z_{d,v,.})}{\sum_{k=1}^{K}[\alpha + \sum_{v=1}^{V} E(z_{d,v,k})]}$$

$$E(\phi_k|\tilde{\beta_k}) = \frac{\beta + \sum_{d=1}^{D} n_{d,v} \times E(z_{d,\cdot,k})}{\sum_{v=1}^{V}(\beta + \sum_{d=1}^{D} n_{d,v} \times E(z_{d,v,k}))}$$

Take home message Z

- NMF is OLS:

$$\hat{Z}_{d,k} = \sum_{v=1}^{V} W_{d,v} Q_{v,k}$$

Take home message Z

- NMF is OLS:

$$\hat{Z}_{d,k} = \sum_{v=1}^{V} W_{d,v} Q_{v,k}$$

- PLSA is MLE:

$$p(z_k|d_i) = \frac{\sum_{v=1}^{V} n(d_i, w_v) p(z_k|d_i, w_v)}{\sum_{k=1}^{K} \sum_{v=1}^{V} n(d_i, w_v) p(z_k|d_i, w_v)}$$

Take home message Z

- NMF is OLS:

$$\hat{Z}_{d,k} = \sum_{v=1}^{V} W_{d,v} Q_{v,k}$$

- PLSA is MLE:

$$p(z_k|d_i) = \frac{\sum_{v=1}^{V} n(d_i, w_v) p(z_k|d_i, w_v)}{\sum_{k=1}^{K} \sum_{v=1}^{V} n(d_i, w_v) p(z_k|d_i, w_v)}$$

- LDA is Bayesian:

$$E(\theta_d|\tilde{\alpha_d}) = \frac{\alpha + \sum_{v=1}^{V} n_{d,v} E(z_{d,v,.})}{\sum_{k=1}^{K} [\alpha + \sum_{v=1}^{V} E(z_{d,v,k})]}$$

Take home message B

- NMF is OLS:

$$\hat{B}_{k,v} = \sum_{d=1}^{D} W_{d,v} P_{k,d}$$

Take home message B

- NMF is OLS:

$$\hat{B}_{k,v} = \sum_{d=1}^{D} W_{d,v} P_{k,d}$$

- PLSA is MLE:

$$p(w_v|z_k) = \frac{\sum_{d=1}^{D} n(d_i, w_v) p(z_k|d_i, w_v)}{\sum_{v=1}^{V} \sum_{d=1}^{D} n(d_i, w_v) p(z_k|d_i, w_v)}$$

Take home message B

- NMF is OLS:

$$\hat{B}_{k,v} = \sum_{d=1}^{D} W_{d,v} P_{k,d}$$

- PLSA is MLE:

$$p(w_v|z_k) = \frac{\sum_{d=1}^{D} n(d_i, w_v) p(z_k|d_i, w_v)}{\sum_{v=1}^{V} \sum_{d=1}^{D} n(d_i, w_v) p(z_k|d_i, w_v)}$$

- LDA is Bayesian:

$$E(\phi_k|\tilde{\beta}_k) = \frac{\beta + \sum_{d=1}^{D} n_{d,v} * E(z_{d,.,k})}{\sum_{v=1}^{V}(\beta + \sum_{d=1}^{D} n_{d,v} * E(z_{d,v,k}))}$$