Topic Models: from Non Negative Matrix Factorization to Latent Dirichlet Allocation

Salfo Bikienga*

September 13, 2018

Abstract

Topic Modeling (TM) is a text data dimension reduction algorithm, akin to factor analysis (FA) or principal component analysis (PCA), widely used for text data analysis (classification, clustering, etc.). Modern TM algorithms such as Latent Dirichlet Allocation (LDA) are probabilistic and complex, impeding their intuitive understanding. However, relating them to Non-Negative Matrix Factorization (NMF), and PCA mitigates this impediment. Indeed, parallel to being analogous to NMF, LDA also emerges from Principal Component Analysis (PCA), both of which are intuitively easy to understand. Therefore, presenting LDA as emerging from NMF and/or PCA provides an intuitive grounding of modern TM algorithms.

1 Introduction

Topic modeling methods are a class of latent variables methods (factor models) applied to text data. Modern topic modeling algorithm originated from principal component analysis (PCA), one

*Email: sbikienga@gmail.com (Please feel free to contact me with any suggestions, corrections or comments.)

of the oldest latent variables methods (Hotelling, 1933). Indeed, Latent Semantic Analysis (LSA), one of the oldest topic modeling algorithm is literally PCA applied to text data (Landauer and Dumais, 1997a). Modern topic modeling algorithms, such as Latent Dirichlet Allocation (LDA) involves numerical approximation of probabilistic words count models, precluding an intuitive understanding of these methods. However, an intuitive understanding of these methods can be developed by demonstrating the rise of such models from traditional models such as PCA, and Non Negative Matrix Factorization (NMF).

An attractive property of PCA is that it does not require a data generating process, making it applicable to all sort of data regardless of how the data generating process. Thanks to this non-parametric property, PCA (particularly the Singular Value (SVD) approach to PCA) was applied to text data, and termed LSA (Landauer and Dumais, 1997a). Yet, the non-parametric property of PCA is also a weakness for inferential studies, which requires probabilistic modeling. Moreover, LSA factors can be difficult to interpret because the factors loadings can take any real value. Since text data are count data, the factor loadings are easier to interpret if they are all non negative. To address these two shortcomings of LSA, Probabilistic Latent Semantic Analysis (PLSA) was proposed (Hofmann, 1999, 2001). PLSA can be seen as a probabilistic NMF algorithm (Paatero and Tapper, 1994; Hubert et al., 2000).

PLSA is prone to over-fitting issues, common with maximum likelihood estimation (MLE) methods, particularly when the dimension is relatively large for the sample size. Thus, Latent Dirichlet Allocation (LDA), a Bayesian approach to PLSA, was proposed to address the over-fitting issue of PLSA (Blei et al., 2003). Bayesian approach differs from MLE approach on the assumption of prior distribution of the parameters to be estimated.

Modern topic modeling algorithms are variant of LDA; and deviate from LDA by the distributional assumption of the parameters' priors (Blei and Lafferty, 2007; Mimno and McCallum, 2012; Roberts et al., 2016); or by the estimation methods (Variational Bayes and Markov Chain Monte Carlo methods) (Blei et al., 2003; Griffiths and Steyvers, 2004; Taddy, 2012; Wang and Blei, 2013).

2 Non Negative Matrix Factorization

Matrix factorization stems from the idea that any matrix can be decomposed into the product of two or more matrices. It turns out that this decomposition can be used to reduce a high dimensional data into a smaller dimension. There are several algorithms for matrix factorization; and the technique has been used mostly for two purposes (Hubert et al., 2000): (1) for solving linear systems (examples include the LU and QR decompositions); (2) For statistical purposes. Among the most popular algorithms used for statistical purposes are: Factor Analysis (FA), Principal Component Analysis (PCA) and/or Latent Semantic Analysis (LSA), Non Negative Matrix Factorization (NMF), Probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA), etc.

2.1 Matrix Factorization

Assume a matrix of data $W_{D\times V}$, representing V variables observed on D individuals. Matrix factorization (MF) postulates that W can be decomposed as follows:

$$W_{D\times V} \simeq Z_{D\times K} B_{K\times V}, \tag{2.1}$$

where $K << V$ is an arbitrary number. An illustrative example may be instructive.

Example 1. Consider the following example data in Table 1. Let $W_{D\times V}$ be a document term matrix, where D represents the number of documents, and V represents the vocabulary list (i.e. the number of unique words). The $W_{d,v}$ element of W represents the count of word v in document d. Consider a matrix W of 6 documents and 5 words. Let the words be: ***college, education, family, health,*** and ***medicaid*** (Table 1). On purpose, rows 1, 3, and 6 have high values for ***college*** and ***education*** and rows 2, 4, and 5 have high values for ***health*** and ***medicaid***.

MF informs that $W_{6,5}$ can be decomposed into the product of two matrices as follows:

Table 1: 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	2	6	1	4	5

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

Note that the similarities between documents observed on the $W_{D,V}$ matrix are also observed on the $Z_{D,K}$ matrix. In fact, $Z_{D\times K}$ explains the variations between the D observations in the original $W_{D\times V}$ dataset. To see why that is the case, observe that eq. 2.1 is similar to a multivariate regression equation; and the underlying assumption of a regression equation is that the features matrix X (represented here as Z) explains the variations observed in the multivariate response Y (represented here as W). Likewise, B is similar to the matrix of coefficients in a multivariate regression model. Moreover, by taking the transpose of eq.2.1 and by considering the new rows to be observations, it can be seen that the B^T matrix explains the observed variations between the V columns of W.

Analyzing reduced dimensional data $Z_{D\times K}$ ($K << V$) presents several advantages. It can alleviate overfitting issues when V is too large compared to D; each row of B can embeds a latent concept that can have practical meaning; for K small enough, graphical tools can be used to explore the data. For example, Fig. 2.1 shows that observations 1, 3, and 6 are close to each other; and

observations 2, 4, 5 are close to each other. A careful observation of the $W_{6,5}$ matrix reveals the same proximity between the observations. However, these proximities are not easily detected with 5 variables.

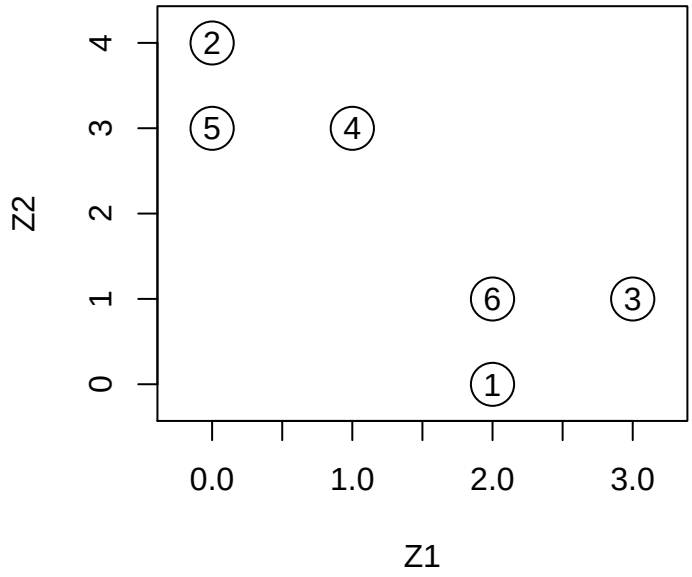

Figure 2.1: Scatterplot of the two dimensional Z variables

For exploratory data analysis, it is essential to know the meaning of the Z variables, since these meanings are the foundation for telling stories from the data. The k^{th} row of B is a list of the total weight of each of the original variables. For our hypothetical example, looking at the third matrix, the first row is dominated by the first two original variables, college and education, suggesting that the first column of Z is an education index (latent concept) variable. Similarly, it is safe to infer that the second column of Z is a health index.

The matrix decomposition in the above example was manually constructed to illustrate the idea of matrix factorization as applied to data analysis. The least squares method is a simple general purpose algorithm for solving Eq. (2.1) for Z and B. In fact, observe that eq. 2.1 is a multivariate linear regression equation, without the error terms; except that Z is not observed.

By defining:

$$W_{D\times V} = Z_{D\times K} B_{K\times V} + \varepsilon_{D\times V}, \tag{2.2}$$

the least squares solution for B in equation 2.2 is given by:

$$\hat{B} = \left(Z^T Z\right)^{-1} Z^T W, \tag{2.3}$$

where it is assumed that $\left(Z^T Z\right)$ is non-singular. If we were solving for Z, the least squares solution is given by:

$$\hat{Z} = W B^T (BB^T)^{-1}, \tag{2.4}$$

where it is assumed that (BB^T) is non-singular. Note that Z, nor B are known. The trick is to initialize Z and solve for B; then use the estimated $\hat{B}$ to solve for a new $\hat{Z}$. Iterate this two step process until convergence of an objective function. A sensible objective function is the euclidean norm (often referred as the L2-norm or Frobenius norm) between W and its predicted values, $\hat{W}$, defined as:

$$\hat{W} = \hat{Z}\hat{B} \tag{2.5}$$

and the objective function is:

$$Q(Z,B) = ||W - ZB||^2$$

where $||\cdot||^2$ is the $L2$-norm.

The optimization problem becomes:

$$[\hat{Z},\hat{B}] = \min_{Z,B} Q(Z,B) \tag{2.6}$$

An R code implementation of this algorithm is presented in Appendix A.1. It should be noted that

problem 2.6 does not have a unique solution. Consider any orthogonal matrix $T_{K\times K}$. If $\hat{Z}$ and $\hat{B}$ are solutions for problem 2.6, then by eq. 2.5, we have $\hat{W} = \hat{Z}\hat{B} = [\hat{Z}T][T^T\hat{B}] = \tilde{Z}\tilde{B}$, and $\tilde{Z}$ and $\tilde{B}$ are also solution of problem 2.6.

It may be informative to re-write the least squared solution expression of B as:

$$\begin{aligned}\hat{B}_{K\times V} &= \left[Z^T Z\right]^{-1} Z^T W \\ &= P_{K\times D} W_{D\times V}\end{aligned}$$

where $P = \left[Z^T Z\right]^{-1} Z^T$ and can be seen as a weight matrix. Then, the scalar elements $\hat{B}_{k,v}$ of the weight matrix B can be written as:

$$\hat{B}_{k,v} = \sum_{d=1}^{D} P_{k,d} W_{d,v} \tag{2.7}$$

From Eq. (2.7), it can be seen that for a given row k of B (which is often referred as topic k in text analytics), the relative weight of the word w_v (noted $\hat{B}_{k,v}$) is the total weighted sum of the word w_v counts in all the documents $W_{1,v}, W_{2,v} \cdots W_{D,v}$, with respective weights $P_{k,1}, P_{k,2}, \cdots, P_{k,D}$.

Likewise, the least squared solution expressions of Z can be re-written as:

$$\begin{aligned}\hat{Z}_{D\times K} &= WB^T \left[BB^T\right]^{-1} \\ &= W_{D\times V} Q_{V\times K}\end{aligned}$$

where $Q = B^T \left[BB^T\right]^{-1}$ and can be seen as a weight matrix. Then, the scalar elements of the reduced dimension data matrix $\hat{Z}_{d,k}$ can be written as:

$$\hat{Z}_{d,k} = \sum_{v=1}^{V} Q_{v,k} W_{d,v} \tag{2.8}$$

From Eq. (2.8), it can be seen that for a given document d, the index value of topic k is a weighted sum of all the words $W_{d,1}, W_{d,2}, \cdots, W_{d,V}$ used in that documents, with respective weights

$Q_{1,k}, Q_{2,k}, \cdots, Q_{V,k}$.

Eq. (2.7) and (2.8) provide an intuitive understanding of topic modeling. Eq. (2.7) says that the relative importance of the word v (W_v) for a given topic k is a weighted sum of the counts of that word in all the documents; and Eq. (2.8) says that the relative importance of a topic in a document is a weighted sum of the counts of all the words used in that document.

It should be noted that despite having non negative values in W, Z and B may have negative values. This poses some interpretational challenges. To see the issue, consider W to be a matrix of words counts, and assume the k^{th} row of B has negative coefficients for some words and positive coefficients for some other words. How would we interpret the relationship between two words with opposing signs? However, if the coefficients are non negative, they have an ordinal meaning. Hence, Non Negative Matrix Factorization (NMF) algorithms were proposed to address this concern (Paatero and Tapper, 1994; Hubert et al., 2000).

This section has presented an intuitive and general idea of topic modeling, by applying NMF algorithm to a hypothetical text data.

2.2 Non Negative Matrix Factorization

The algorithm presented above can be labeled constraint-less matrix factorization algorithm. There is a plethora of MF algorithms with varying constraints and/or assumptions.

- PCA (Hotelling, 1933) and FA (Thurstone, 1935) impose orthogonality constraint between the $Z_{\cdot k}$ vectors and between the $B_{k\cdot}$ vectors.
- Trapezoid MF imposes $Z_{i,j} = 0$ for $j > i$, $B_{i,j} = 0$ for $j < i$ (LU factorization is an example).
- We can think of these methods as special cases of the generic algorithm presented in section 2.1.

Another form of constraint is to impose $Z_{i,j} \geq 0$, $B_{i,j} \geq 0$ for all i, j. Imposing such constraint leads to the notion of Positive Matrix Factorization (PMF), generally referred as Non Negative Matrix

Factorization; and originally termed Non-Negative Factor Model (Paatero and Tapper, 1994).

Let's write the MF as:

$$W_{D\times V} \simeq Z^{nmf} B^{nmf} \tag{2.9}$$

where the subscript ***nmf*** emphasizes non negativity of the elements of Z and B.

In practice, Z^{nmf} and B^{nmf} are estimated by alternating least squares where the negative values $Z^{nmf}_{d,k}$ and $B^{nmf}_{k,v}$ are set to zero at each iteration (Paatero and Tapper, 1994). A simple R implementation of this algorithm is presented in appendix A.2. Berry et al. (2007) presents a survey of different algorithms and applications of NMF. Eldén (2007, chap.9) presents a textbook treatment of NMF.

The NMF can be normalized so that Z and B can be interpreted as probability distributions (Gillis, 2017). To do so, let D_W and D_B be two diagonal matrices, with diagonal elements the rows sums of W and B respectively (i.e. $D_{W_{d,d}} = \sum_{v=1}^{V} W_{d,v}$; $D_{B_{k,k}} = \sum_{v=1}^{V} B_{k,v}$). Further assume $W = ZB$. Then, we can write:

$$\begin{aligned} D_W^{-1} W &= \left[D_W^{-1} Z D_B\right]\left[D_B^{-1} B\right] \\ &\Longleftrightarrow \\ W^* &= Z^* B^* \end{aligned} \tag{2.10}$$

where $W^* = D_W^{-1} W$, $Z^* = D_W^{-1} Z D_B$, and $B^* = D_B^{-1} B$.

It is easy to see that the rows sums of B^* are all $1s$; since each $B^*_{k,v} = \frac{B_{k,v}}{\sum_{v=1}^{V} B_{k,v}}$ by definition, $\sum_{v=1}^{V} B^*_{k,v} = \sum_{v=1}^{V} \frac{B_{k,v}}{\sum_{v=1}^{V} B_{k,v}} = \frac{1}{\sum_{v=1}^{V} B_{k,v}} \sum_{v=1}^{V} B_{k,v} = 1$. Thus, $0 \leq B^*_{k,v} \leq 1$, and $\sum_{v=1}^{V} B^*_{k,v} = 1$.

To see why the rows sums of Z^* are all $1s$, observe that for each row i of W^*, $\sum_{v=1}^{V} W^*_{d,v} = 1$ by definition. So, we have

$$
\begin{aligned}
1 &= \sum_{v=1}^{V} W^*_{d,v} \\
&= \sum_{v=1}^{V} \left[\sum_{k=1}^{K} Z^*_{d,k} B^*_{k,v} \right] \\
&= \sum_{k=1}^{K} \left[Z^*_{d,k} \sum_{v=1}^{V} B^*_{k,v} \right] \\
&= \sum_{k=1}^{K} \left[Z^*_{d,k} * 1 \right] \\
&= \sum_{k=1}^{K} Z^*_{d,k}
\end{aligned}
\tag{2.11}
$$

Thus, $0 \leq Z_{d,k} \leq 1$, and $\sum_{k=1}^{K} Z_{d,k} = 1$. Therefore, each row d of Z^* and each row k of B^* is a probability distribution (by the definition of discrete probability distribution).

Even though the elements of Z and B can be interpreted as probabilities, this interpretation is ad hoc. A formal use of probability distribution for NMF in text analytics is presented in section 5.

3 Principal Component Analysis

A parallel application of matrix factorization algorithms in text mining consists of the application of PCA to text data. PCA is one of the oldest dimension reduction algorithm (Hotelling, 1933). It consists of re-expressing the observed data $X_{n\times p}$ into $Z_{n\times K}$ with the goal of removing redundancy (defined as correlation between variables) in the original data X, while preserving most of the variations in X. n, p, and K represents the number of observations, the number of variables in the original dataset, and the number of the variables of the reduced dataset, respectively. Geometrically, PCA consists of orthogonally projecting the p dimensional data X into a K sub-dimensional space.

Let P be an orthogonal projection matrix. We can define Z as follows:

$$Z_{n\times p} = X_{n\times p} P_{p\times p} \tag{3.1}$$

Define:

$$\begin{aligned} C_Z &= \frac{1}{n-1} Z^T Z \\ &= \frac{1}{n-1} \left[P^T X^T X P \right] \\ &= P^T \left[\frac{1}{n-1} X^T X \right] P \\ &= P^T C_X P \end{aligned} \tag{3.2}$$

where C_Z and C_X are the covariance matrices of Z and X, respectively.

Theorem. *If A is symmetric, there is an orthogonal matrix E such that $A = EDE^T$, where D is a diagonal matrix (Gentle, 2017, p.154).*

Using this theorem, and setting $P = E$, we have:

$$\begin{aligned} C_Z &= P^T C_X P \\ &= E^T \left[EDE^T \right] E \\ &= E^T EDE^T E \\ &= D \end{aligned} \tag{3.3}$$

Thus, the covariance matrix of Z is indeed a diagonal matrix, that is, the Z variables are not correlated. Note that by orthogonality of E, $E^T E = EE^T = I$.

Definition. The total variance is the trace of the covariance matrix.

We observe from eq. 3.3 that:

$$
\begin{aligned}
tr(C_Z) &= tr(D) \\
&= tr(P^T C_X P) \\
&= tr(E^T C_X E) \\
&= tr(E E^T C_X) \\
&= tr(C_X)
\end{aligned}
\tag{3.4}
$$

From eq. 3.3 and 3.4, it is confirmed that the covariance of Z is diagonal, and Z preserves the total variance of the original dataset X.

PCA as a dimension reduction methods arises from observing that D is a rank ordered matrix, that is, the diagonal elements of the D matrix are ordered in decreasing order. Consequently, the hope is that there is a $K << p$, such that $\sum_{k=1}^{K} d_{k,k} \simeq tr(C_X)$; in which case,

$$
Z_{n\times K} \simeq X_{n\times p} E_{p\times K} \tag{3.5}
$$

approximates the original dataset $X_{n\times p}$.

We can approximately retrieve $X_{n\times p}$ from eq. 3.5:

$$
\begin{aligned}
Z_{n\times K} E^T_{K\times p} &\simeq X_{n\times p} E_{p\times K} E^T_{K\times p} \\
&\Longleftrightarrow \\
X_{n\times p} &\simeq Z_{n\times K} E^T_{K\times p}
\end{aligned}
\tag{3.6}
$$

Eq. 3.6 shows that PCA is a matrix factorization algorithm. Unlike the matrix factorization algorithms presented in Section 2, PCA is an orthogonal matrix factorization algorithm (by the above theorem), and does not impose non negativity of the factorized matrices.

The above theorem is known as eigen decomposition (or spectral decomposition). The Singular

Value Decomposition (SVD) algorithm is a more general solution to the PCA problem (Shlens, 2014). SVD generalizes the notion of eigenvectors from square matrices (such as the covariance matrix) to any kind of matrix (Murphy, 2012, p.394).

By SVD, any real $n \times p$ matrix X can be decomposed as follows:

$$X_{n\times p} = U_{n\times n} S_{n\times p} V^T_{p\times p} \tag{3.7}$$

where U and V are orthonormal matrices, that is $U^TU = UU^T = I_n$, $V^TV = VV^T = I_p$. S is a diagonal matrix containing the $r = min(n,p)$ singular values $\sigma_k \geq 0$ on the main diagonal, with 0s filling the rest of the matrix.

If $X_{n\times p}$ are zero means p variables, the covariance matrix:

$$\begin{aligned} C_X &= \frac{1}{n-1} X^T X \\ &= \frac{1}{n-1} VSU^T USV^T \\ &= \frac{1}{n-1} VS^2V^T \\ &= VDV^T \end{aligned} \tag{3.8}$$

where $D = \frac{1}{N-1}S^2$. Eq. 3.8 reveals that the cross product of the SVD of X yields the eigen-decomposition of the covariance matrix of X. But, the eigen-decomposition of C_X is the PCA solution; hence, the equivalence of PCA and SVD.

Assuming $n > p$, eq. 3.7 can be re-written as:

$$X_{n\times p} = U_{n\times p} S_{p\times p} V^T_{p\times p} \tag{3.9}$$

Moreover, if there is a K such that $\sigma_{K+i} \simeq 0$, for $i = 1, 2, \cdots, P-K$, we can approximate $X_{n\times p}$, by $\hat{X}_{n\times p}$, where:

$$\hat{X}_{n\times p} = U_{n\times K} S_{K\times K} V^T_{K\times p} \tag{3.10}$$

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

$$\hat{X} \simeq ZB \tag{3.11}$$

Again, PCA (whether it is solved by eigen-decomposition or SVD) is a matrix factorization algorithm, meant to transformed a high dimensional data $X_{n\times p}$ into a lower dimensional data $Z_{n\times K}$ (where $K \leq p$). The lower dimensional data preserves most of the total variations in $X_{n\times p}$, and is orthogonal.

Modern TM algorithm derives directly from PCA. The main goal of this section was to present PCA, and show the similarity between PCA and MF, which is easy to grasp, intuitively.

4 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is an application of SVD to a matrix of words counts (Landauer and Dumais, 1997b). As such, LSA is exactly another application of PCA.

To fix ideas, consider the example dataset provided in example 1. Applying PCA (eigen-decomposition) to the example dataset, it appears that the first three components captures all the variations in $W_{6\times 5}$ (see standard deviation values in the output); Moreover, about 92.6% of the total variation in W is explained by the first two components, suggesting that the five-dimensional data can be effectively reduced into a two-dimensional data.

```
W = scale(W, center = TRUE, scale = FALSE) # to standardize the W variables
pca_W = princomp(W) # princomp() is one of the R software function for PCA
pca_W$sdev

##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## 5.28e+00 2.11e+00 5.88e-01 6.57e-09 0.00e+00
```

The term PC-scores is often used to refer to the Z values.

```
pca_scores = data.frame(pca_W$scores)
```

Table 2: PC scores or Z values

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
document.1	-6.580	2.580	0.637	0	0
document.2	7.350	-0.511	0.017	-0	-0
document.3	-5.370	-3.530	0.243	-0	0
document.4	3.110	-1.520	0.093	-0	-0
document.5	4.630	2.040	0.252	0	-0
document.6	-3.150	0.929	-1.240	0	0

The B matrix from $W_{D\times V} \simeq Z_{D\times K} B_{K\times V}$ is known as the loadings matrix and is used to interpret the meaning of the new Z variables. Table 2 shows the PC scores or the Z values. Observe that only the first two columns have sizable scores. Since the first two components are deemed worthwhile interpreting, the meaning of the two Z variables can be inferred from the first two rows of the loadings. The first row of Table 3 suggests that the first component is a contrast between education and health (the education variables have high negative scores while the health variables have high positive scores). The second component does not have a clear interpretation since the education and health related variables have all high scores. Another thing to note is that despite words counts being positive, the PC scores and their loadings have negative and positive values. That poses interpretational challenges; for example, referring to component 3 in Table 3, it is not clear what meaning can be inferred from the opposite loadings of education and college.

```
pca_loadings = t(pca_W$loadings[,]) # extracting the factor loadings
```

Next, let's check how the SVD solution compares to the regular PCA (eigen-decomposition) solution. Comparing Table 2 and Table 4 suggests that the PC scores are identical whether the solution is obtained by eigen-decomposition or SVD.

Table 3: PC loadings

	college	education	family	health	medicaid
Comp.1	-0.350	-0.582	0.264	0.334	0.598
Comp.2	-0.449	-0.570	-0.115	-0.419	-0.534
Comp.3	0.822	-0.560	0.050	-0.087	-0.037
Comp.4	0	0.078	-0.105	-0.808	0.574
Comp.5	0	-0.131	-0.951	0.228	0.165

```
svd_W = svd(W) # Singular Value Decomposition applied to the data

U = svd_W$u

S = diag(svd_W$d)

Z_svd = svd_scores = U %*%S
```

Table 4: PC scores obtained by SVD

-6.580	2.580	0.637	-0	0
7.350	-0.511	0.017	-0	0
-5.370	-3.530	0.243	0	-0
3.110	-1.520	0.093	-0	-0
4.630	2.040	0.252	0	-0
-3.150	0.929	-1.240	-0	-0

Last, let's perform a LSA on the same data. Looking at the 'lsa()' output ('lsa()' is the R function to perform LSA), it can be noted that the output is identical to the svd output (multiplying the first matrix–$tk–of the output below with the diagonal matrix of $sk vector yields the same score as the pc-scores above).

This exercise has shown that PCA, SVD, and LSA are the same methods. Thus, it is clear that LSA, the oldest topic modeling algorithm is simply a PCA algorithm (SVD) applied to text data.

```
library(lsa) # Needed to access the lsa function

lsa_W = lsa(W)

lsa_W # print the output
```

```
## $tk
##               [,1]    [,2]    [,3]    [,4]    [,5]
## document.1 -0.509  0.4999  0.4418 -0.3261  0.1869
## document.2  0.569 -0.0988  0.0120 -0.0647  0.6561
## document.3 -0.415 -0.6826  0.1686  0.3528 -0.0450
## document.4  0.241 -0.2934  0.0642 -0.7183 -0.5170
## document.5  0.358  0.3951  0.1747  0.4962 -0.5129
## document.6 -0.244  0.1797 -0.8612 -0.0520 -0.0464
##
## $dk
##             [,1]   [,2]    [,3]    [,4]    [,5]
## college   -0.350 -0.449  0.8218  0.0000  0.0000
## education -0.582 -0.570 -0.5597  0.0562  0.1418
## family     0.264 -0.115  0.0499 -0.2556  0.9216
## health     0.334 -0.419 -0.0868 -0.7616 -0.3542
## medicaid   0.598 -0.534 -0.0369  0.5929 -0.0711
##
## $sk
## [1] 1.29e+01 5.17e+00 1.44e+00 4.52e-16 1.34e-16
##
## attr(,"class")
## [1] "LSAspace"

# lsa_W$tk%*%diag(lsa_W$sk) # to check equality of pc-scores
```

Fig. 4.1 illustrates the usefulness of dimension reduction methods. The example data has five variables (words); it is not trivial to detect similarities or differences between the observations (documents) by considering all five variables. However, by PCA, the five variables are collapsed

into two composite variables, which we can easily explore with a scatterplot (or biplot). Fig. 4.1 shows that education related documents (documents 1, 3, and 6) are located to the left of the origin of the first component, and the health related documents (documents 2, 4, and 5) are located to the right of the first component. Thus we can surmise a similarity between documents 1, 3, and 6, and a similarity between documents 2, 4, and 5.

```
biplot(pca_W, cex = 0.5, xlim = c(-0.8, 0.8))
```

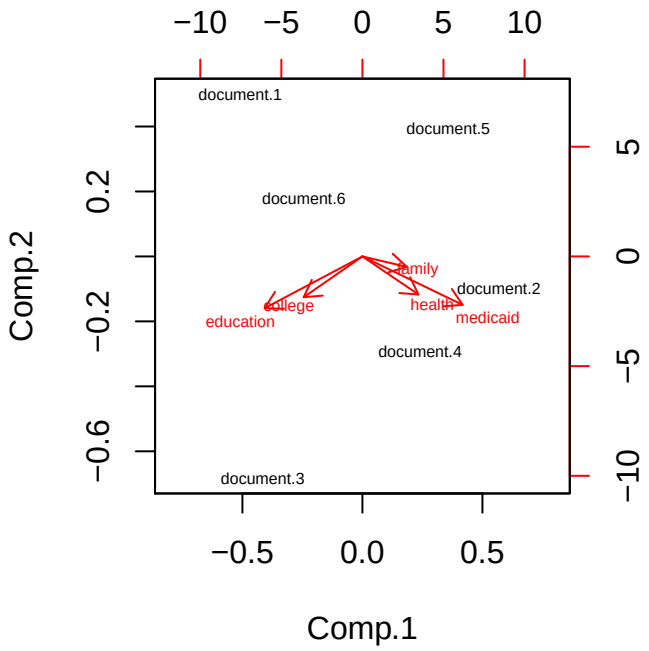

Figure 4.1: Biplot of the PCA results

PCA simply uses matrix algebra to compress information without regards to the characteristics of the data. For instance, it does not take into consideration that the variables are positive or count data. That raises interpretational concerns. Referring to Fig. 4.1, there is nothing to say about education variables being positioned to the left, and health variables being positioned to the right. All the figure tells is that there is a contrast between these variables. To see why this shortcoming is important, let's consider the first dimension of Fig. 4.1; based on the first axis, we can only say

that, for example documents 5 and 6 are in opposite directions. There is no information related to the relative importance of these documents with respect to the first latent variable. If we could force the axis to start from zero, then the relative position of a document on the axis informs on the relative importance of the document with respect to the particular latent variable that the first axis represents; that is, the axis' numbers have an ordinal interpretation. A solution to the lack of ordinal interpretation of the coefficients is to use the Non Negative Matrix Factorization (NMF) method presented in Section 2.2, or to use a probabilistic model.

5 Probabilistic Latent Semantic Analysis

To fully understand topic modeling, we have performed two tasks: (1) we have shown that LSA, the oldest of topic modeling algorithms is identical to PCA; (2) these data dimension reduction methods are all special cases of matrix factorization algorithms. In statistics, these methods are known as latent variable high dimensional data reduction methods. However, they all suffer from the lack of a probability model, since they do not require a distributional assumption of the data. This shortcoming is not an issue for exploratory data analysis. Though they can reduce the dimension of the data well, they are not suited for inferential studies. Inferential studies require model assumptions; by assuming a data generating process (probability model), inference statistics relies on the parameters estimates of the model to generalize from a sample data to a population. Probabilistic Latent Semantic Analysis (PLSA) was proposed to supplement the lack of distributional assumption of LSA (Hofmann, 1999, 2001). In that sense, PLSA is a "statistical view of LSA" (Hofmann, 1999, p.289).

5.1 PLSA: Model Specification

Table 5: Define terms

D	Number of documents
d_i	The i^{th} document
V	Number of unique words
w_v	The v^{th} word
n_{d_i,w_v}	The number of words w_v in document d_i
Z	Topic identity or topic label

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

We can write:

$$\begin{aligned} p(w_v|d_i) &= \sum_{z\in Z} p(w_v, z|d_i) \\ &= \sum_{z\in Z} p(w_v|z, d_i)p(z|d_i) \\ &= \sum_{z\in Z} p(w_v|z)p(z|d_i) \end{aligned} \tag{5.1}$$

Where z is a categorical hidden (or latent) variable taking values $1, 2, \cdots, K$. z indicates the topic identity (or topic number). The second line of Eq. 5.1 derives from the Bayes theorem, and the third line derives from the assumption that conditional on z, w_v and d_i are independent.

A document is a collection of $N_{d_i} = \sum_v^V n_{d_i,w_v}$ words, assumed independent; thus, given a document, the joint probability of observing the words $w_1, w_2, \cdots, w_V$ is:

$$p(w_1, w_2, \cdots, w_V|d_i) = \prod_{v=1}^{V} p(w_v|d_i)^{n(d_i,w_v)} \tag{5.2}$$

Eq. 5.2 is a multinational distribution without the normalizing constant. It may be helpful here to think of a V sided die of unequal probabilities (multinouilly or categorical distribution). Observing a single word is equivalent to rolling the die once. Observing the N_{d_i} words of document d_i is equivalent to rolling the V sided die N_{d_i} times, independently.

Assuming independence of the documents, the joint probability (or the joint likelihood) of observing the corpus (collection of D documents) is:

$$p(W|D) = \prod_{d=1}^{D} \prod_{v=1}^{V} p(w_v|d_i)^{n(d_i,w_v)} \tag{5.3}$$

By the third line of Eq. 5.1, Eq. 5.3 becomes:

$$L(\theta|W) = \prod_{d=1}^{D} \prod_{v=1}^{V} \left(\sum_{z \in Z} p(z|d_i) p(w_v|z) \right)^{n(d_i,w_v)} \tag{5.4}$$

Taking the log of eq. 5.4 gives:

$$\mathscr{L}(\theta|W) = \sum_{d=1}^{D} \sum_{v=1}^{V} n(d_i, w_v) log \left(\sum_{z \in Z} p(w_v|z) p(z|d_i) \right) \tag{5.5}$$

The estimation goal is to find $p(w_v|z)$ and $p(z|d_i)$ to maximize $\mathscr{L}(\theta|W)$. However, due to the sum operator inside the log, there is no closed form solution for this problem. Thus, the traditional maximum likelihood estimation (MLE) method does not apply here. An alternative is to use the Expectation Maximization (EM) algorithm. It is a numerical method which aims at approximating the log likelihood (5.5).

It can be shown (see Appendix (B)) that the approximation solution of (5.5) consists of alternatively solving:

$$p(z_k|d_i, w_v) = \frac{p(w_v|z_k) p(z_k|d_i)}{\sum_{l=1}^{K} p(w_v|z_l) p(z_l|d_i)} \tag{5.6}$$

in the E-step, and

$$p(w_v|z_k) = \frac{\sum_{d=1}^{D} n(d_i, w_v) p(z_k|d_i, w_v)}{\sum_{v=1}^{V} \sum_{d=1}^{D} n(d_i, w_v) p(z_k|d_i, w_v)} \tag{5.7}$$

$$p(z_k|d_i) = \frac{\sum_{v=1}^{V} n(d_i, w_v) p(z_k|d_i, w_v)}{\sum_{k=1}^{K} \sum_{v=1}^{V} n(d_i, w_v) p(z_k|d_i, w_v)} \tag{5.8}$$

in the M-step, until convergence of some objective function defined as,

$$Q(\theta) = \sum_{d=1}^{D} \sum_{v=1}^{V} n(w_v, d_i) \sum_{k=1}^{K} p(z_k|w_v, d_i) log\left(p(w_v|z_k)p(z_k|d_i)\right) \tag{5.9}$$

where θ is a short hand notation for the parameter to be estimated.

In practice, initial values are provided (or assumed) for $p(w_v|z_k)$ and $p(z_k|d_i)$ to compute $p(z_k|w_v, d_i)$ in the E-step; then the computed $p(z_k|w_v, d_i)$ is used to compute new values for $p(w_v|z_k)$ and $p(z_k|d_i)$ in the M-step. E-step, M-step are computed alternatively until convergence of Eq. (5.9).

5.2 Relation between LSA and PLSA

PLSA is a model based approach to LSA, and as such the two methods are fundamentally linked (Hofmann, 1999). To see this link, let's write the joint distribution between the word w_v and the document d_i as:

$$\begin{aligned} p(w_v, d_i) &= \sum_{Z} p(z)p(w_v|z)p(d_i|z) \\ &= \sum_{z_k=1}^{K} p(d_i|z_k)p(z_k)p(w_v|z_k) \end{aligned} \tag{5.10}$$

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:

$$\begin{aligned} [p(w_v, d_i)]_{D\times V} &= \sum_{z_k=1}^{K} p(d_i|z_k)p(z_k)p(w_v|z_k) \\ &= [p(d_i|z_k)]_{D\times K} [p(z_k)]_{K\times K} [p(w_v|z_k)]_{K\times V} \\ &= USV^T \end{aligned} \tag{5.11}$$

5.3 On the equivalence of PLSA and NMF

Eq. (2.10) reveals that the NMF can be re-expressed in probabilistic form, where the Z and B can be interpreted as probability distributions. That probabilistic interpretation of NMF approach is ad hoc, and is done mostly for interpretational purpose. PLSA is a model based approach with the probabilistic interpretation of the parameters being inherent to the model definition. Gaussier and Goutte (2005) formally show the equivalence between PLSA and NMF.

To show the equivalence between the two methods, define:

$$Z^{plsa} = [p(d_i|z_k)]_{D\times K}$$

and

$$B^{plsa} = [p(z_k)p(w_v|z_k)]_{K\times V}$$

Then, the matrix $\tilde{W}$ of the joint distribution of w_v, d_i of the D documents by V words can be written as:

$$\begin{aligned}[p(w_v,d_i)]_{D\times V} &= [p(d_i|z_k)]_{D\times K}\,[p(z_k)p(w_v|z_k)]_{K\times V}\\ &\Rightarrow\\ \tilde{W} &= Z^{plsa}B^{plsa}\end{aligned}$$

Thus PLSA is a NMF.

Conversely, we can show that NMF is a PLSA. To do so, consider Eq. (2.9). Let's re-scale W into $\bar{W}$ in such a way that $\sum_{d_i,v}\bar{W}_{d_i,v} = 1$. Then $\bar{W}$ can be seen as a matrix of joint distributions of w_v and d_i; Eq. (2.9) becomes:

$$\bar{W}_{D\times V} = Z^{nnf}_{D\times K}B^{nmf}_{K\times V}$$

Let's further introduce D_Z and D_B, two $K\times K$ diagonal scaling matrices such that $D_{Z_{k\times k}} = \sum_{i=1}^{D}\bar{Z}_{d_i,k}$, and $D_{B_{k\times k}} = \sum_{v=1}^{V}\bar{B}_{k,v}$. Then

$$\begin{aligned}\bar{W}_{D\times V} &= Z^{nmf}_{D\times K} D_Z^{-1} D_Z D_B D_B^{-1} B^{nmf}_{K\times B} \\ &= \left[Z^{nmf}_{D\times K} D_Z^{-1}\right] [D_Z D_B] \left[D_B^{-1} B^{nmf}_{K\times B}\right] \\ &= [p(d_i|z_k)]_{D\times K} [p(z_k)]_{K\times K} [p(w_v|z_k)]_{K\times V}\end{aligned}$$

Thus, NMF is a PLSA.

Consequently NMF and PLSA are equivalent.

6 Latent Dirichlet Allocation

The idea of topic modeling is generally associated with PLSA and LDA, and their subsequent variants. In terms of evolution of ideas, LDA is a Bayesian approach to PLSA, and was proposed by Blei et al. (2003). From our model specification of PLSA in Section 5.1, let's define:

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

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

Where $\theta_{d,k}$ is the probability (or proportion) of observing topic k in document d; and $\phi_{k,v}$ is the probability of observing the word w_v given topic k.

For a given document d $\theta_d = [\theta_{d,1}, \theta_{d,2}, \cdots, \theta_{d,K}]$ is a K vector, and for a given topic k, $\phi_k = [\phi_{k,1}, \phi_{k,2}, \cdots, \phi_{k,V}]$ is a V vector, where K is the number of topics in each document, and V is the number of unique words in the corpus (collection of D documents). For PLSA, the θs are parameters to be estimated. LDA assumes the θs and ϕs to be random parameters with prior Dirichlet distributions.

$$\theta_d \sim Dirichlet_K(\alpha)$$

$$\phi_k \sim Dirichlet_V(\beta_k)$$

By assumption, the K vector α is the same for all documents; and each ϕ_k has its own V vector β_k parameters.

Under these assumptions, PLSA has two major shortcomings that LDA solves:

1. The number of parameters to estimate is linear in the number of documents in PLSA. To see this, recall that $p(w_v|d_i) = \sum_{k=1}^{K} p(w_v|z_k)p(z_k|d_i)$ or $[p(w_v|d_i)]_{D\times V} = [p(z_k|d_i)]_{D\times K}[p(w_v|z_k)]_{K\times V}$. From this equation, we see that the number of parameters to estimate is $D\times K + K\times V = K(D+V)$. However, the parameters to estimate for LDA are the K vector α and the $K\times V$ matrix $\beta_{K\times V}$, that is LDA estimates $K + K\times V$ parameters α and β. Given that PLSA has significantly more parameters to estimate than LDA, it is more susceptible to over-fitting issues.

2. Contrary to PLSA, LDA is a fully Bayesian approach, which provides a straightforward way to make inferences about documents not previously seen in the training data.

Despite the shortcomings observed with PLSA, Lu et al. (2011) shows that there is no clear answer to which of the two methods perform better for regular text mining task. However, LDA seems to be better at tasks based on the reduced dimensional representation of the text, θ, (provided that the prior parameters α are optimally chosen).

Girolami and Kabán (2003) shows that by setting $\alpha = 1$, PLSA is a special case of LDA.

Next, let's provide a mathematical exposition of LDA and present the data generative process assumed to estimate the posterior distributions of θ and ϕ. LDA is a model that represents documents as being generated by a random mixture over latent variables called topics (Blei et al., 2003). A topic is defined as a distribution over words. For a given corpus of D documents each of length N_d, the generative process for LDA is defined as follows:

1. For each topic k, draw a vector distribution of words ϕ_k; where $\phi_k \sim Dirichlet(\beta_k)$ with $k = \{1,2,...K\}$

2. For each document d:

   (a) Draw a vector of topic proportions θ_d, where $\theta_d \sim Dirichlet(\alpha)$

   (b) For each word i

      i. Draw a topic assignment $z_{d,n}$, where $z_{d,n} \sim multinomial(\theta_d)$ with $z_{d,n} \in \{1,2,...,K\}$

      ii. Draw a word $w_{d,v}$, where $w_{d,v} \sim multinomial(\phi_{k=z_{d,n}})$ with $w_{d,v} \in \{1,2,...,V\}$

Note: Only the words w are observed.

This formal definition of LDA can be difficult to understand. An informal explanation can be helpful:

Point 1 can be thought of having K boxes containing the same set of words with varying proportions. for example, if the vector ϕ_1 (i.e box 1) is about education, we would expect relatively more words about education and fewer words referring to other concepts (i.e $\phi_{1,education}$,$\phi_{1,school}$, $\phi_{1,college}$, $\cdots$ are relatively high compared to, say $\phi_{1,crime}$, $\phi_{1,travel}$, $\phi_{1,bank}$). Likewise, if the vector ϕ_2 (i.e box 2) is about economy, we would expect relatively more words about economy and fewer words referring to other concepts ((i.e $\phi_{2,economy}$,$\phi_{2,business}$, $\phi_{2,growth}$, $\cdots$ are relatively high compared to, say $\phi_{2,sport}$, $\phi_{2,travel}$, $\phi_{2,police}$). And so on for ϕ_3 (box 3), $\cdots$, ϕ_K (box K).

Point 2 says that for a given a document d, to generate a word $w_{d,n}$, we have to decide first which box the word $w_{d,n}$ should be drawn from. Thus, we first draw a box ID $z_{d,n} = k$. But, the likelihood of $z_{d,n}$ being 1, or 2, or $\cdots$ or K is the likelihood of topic k in the document d, $\theta_{d,k} = p(z_{d,n} = k|\theta)$. So, for a given document, a topic identity $z_{d,n} = k$ is drawn $p(z_{d,n} = k|\theta_d)$, θ_d being the topics proportions in document d, then given that the topic of interest is $z_{d,n} = k$, we draw the words $w_{d,n} = v$ from the box $\phi_{k=z_{d,n}}$ ($p(w_{d,n} = v|\phi_{k=z_{d,n}})$).

The inferential goal is to find the parameters θ and ϕ with the highest likelihood to have generated the observed words w. The above generative process allows us to construct an explicit closed

form expression for the joint likelihood of the observed and hidden variables. Markov Chain Monte Carlo (MCMC), and Variational Bayes methods can then be used to estimate the posterior distribution of θ and ϕ (that is $\alpha, \beta | w$) (Blei et al., 2003; Griffiths and Steyvers, 2004; Blei, 2012). The following is the variational Bayes derivation of the posterior distribution of the θs and ϕs.

A topic ϕ_k is a distribution over V unique words, each having a proportion $\phi_{k,v}$; i.e $\phi_{k,v}$ is the relative importance of the word v for the definition (or interpretation) of the topic k. It is assumed, for simplicity, that:

$$\phi_k \sim Dirichlet_V(\beta)$$

That is, the vectors $\beta_1, \beta_2, \cdots, \beta_K = \beta$, so

$$p(\phi_k|\beta) = \frac{1}{B(\beta)} \prod_{v=1}^{V} \phi_{k,v}^{\beta-1}$$

Where $B(\beta) = \frac{\prod_{v=1}^{V} \Gamma(\beta)}{\Gamma(\sum_{v=1}^{V} \beta)}$. Since we have K independent topics (by assumption),

$$p(\phi|\beta) = \prod_{k=1}^{K} \frac{1}{B(\beta)} \prod_{v=1}^{V} \phi_{k,v}^{\beta-1} \tag{6.1}$$

A document d is a distribution over K topics, each having a proportion $\theta_{d,k}$, i.e. $\theta_{d,k}$ is the relative importance of the topic k, in the document d. We assume:

$$\theta_d \sim Dirichlet_K(\alpha)$$

That is:

$$p(\theta_d|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k-1}$$

And since we have D independent documents (by assumption),

$$p(\theta|\alpha) = \prod_{d=1}^{D} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha-1} \tag{6.2}$$

Let z be the latent topic assignment variable, i.e. the random variable $z_{d,n}$ assigns the n^{th} word $w_{d,n}$, from document d, to the topic k. $z_{d,n}$ is a vector of zeros and 1 at the k^{th} position ($z_{d,n} = [0,0,...1,0,..]$). Define $z_{d,n,k} = I(z_{d,n} = k)$ where I is an indicator function that assigns 1 to the random variable $z_{d,n}$ when $z_{d,n}$ is the topic k, and 0 otherwise.We assume:

$$z_{d,n} \sim Multinomial_K(1, \theta_d)$$

That is:

$$p(z_{d,n,k}|\theta_d) = \theta_{d,k} = \prod_{k=1}^{K} \theta_{d,k}^{z_{d,n,k}}$$

A document d has N_d independent words, and since we assume D independent documents, we have:

$$\begin{aligned} p(z|\theta) &= \prod_{d=1}^{D} \prod_{n=1}^{N_d} \prod_{k=1}^{K} \theta_{d,k}^{z_{d,n,k}} \\ &= \prod_{d=1}^{D} \prod_{k=1}^{K} \prod_{n=1}^{N_d} \theta_{d,k}^{z_{d,n,k}} \\ &= \prod_{d=1}^{D} \prod_{k=1}^{K} \prod_{v=1}^{V} \theta_{d,k}^{n_{d,v} \times z_{d,v,k}} \end{aligned} \tag{6.3}$$

where $z_{d,v,k} = I(z_{d,n} = k)I(w_{d,n} = v)$; $n_{d,v}$ is the count of the word v in document d.

The word $w_{d,n}$ is drawn from the topic's words distribution ϕ_k:

$$w_{d,n}|\phi_{k=z_{d,n,k}} \sim Multinomial(1, \phi_{k=z_{d,n}})$$

$$p(w_{d,n} = v|\phi_{k=z_{d,n}}) = \phi_{k,v}$$
$$= \prod_{v=1}^{V}\prod_{k=1}^{K}\phi_{k,v}^{w_{d,n,v}\times z_{d,n,k}}$$

$w_{d,n}$ is a vector of zeros and 1 at the v^{th} position. Define $w_{d,n,v} = I(w_{d,n} = v)$ where I is an indicator function that assigns 1 to the random variable $w_{d,n}$ when $w_{d,n}$ is the word v, and 0 otherwise.

There are D independent documents, each having N_d independent words, so:

$$p(w|\phi) = \prod_{d=1}^{D}\prod_{n=1}^{N_d}\prod_{v=1}^{V}\prod_{k=1}^{K}\phi_{k,v}^{w_{d,n,v}\times z_{d,n,k}} \tag{6.4}$$
$$= \prod_{d=1}^{D}\prod_{v=1}^{V}\prod_{k=1}^{K}\phi_{k,v}^{n_{d,v}\times z_{d,v,k}}$$

The joint distribution of the observed words w and unobserved (or hidden variables) θ, z, and ϕ is given by:

$$P(w,z,\theta,\phi|\alpha,\beta) = p(\theta|\alpha)p(z|\theta)p(w|\phi,z)p(\phi|\beta)$$

The goal is to get the posterior distribution of the unobserved variables:

$$p(z,\theta,\phi|w,\alpha,\beta) = \frac{P(w,z,\theta,\phi|\alpha,\beta)}{\int\int\sum_z P(w,z,\theta,\phi|\alpha,\beta)d\theta d\phi}$$

$\int\int\sum_z P(w,z,\theta,\phi|\alpha,\beta)d\theta d\phi$ is intractable, so approximation methods are used to approximate the posterior distribution. The seminal paper of LDA (Blei et al., 2003) uses the Mean Field Variational Bayes to approximate the posteriors distribution. A Markov Chain Monte Carlo (MCMC) estimation method for LDA was proposed by Griffiths and Steyvers (2004). The mean field variational

inference assumes independence of the posterior distributions as follow:

$$p(z,\theta,\phi|w,\alpha,\beta) \simeq q(z,\theta,\phi|\cdot) = q(z|\cdot)q(\theta|\cdot)q(\phi|\cdot)$$

where the dot $(\cdot)$ is a place holder for the unknown posterior parameter.

From Bishop (2006, p.466), the posteriors distributions:

$$q^*(z|\cdot) \propto exp\left\{E_{\theta,\phi}\left[log(p(z|\theta)) + log(p(w|\phi,z))\right]\right\} \tag{6.5}$$

$$q^*(\theta|\cdot) \propto exp\left\{E_{z,\phi}\left[log(p(\theta|\alpha)) + log(p(z|\theta))\right]\right\} \tag{6.6}$$

$$q^*(\phi|\cdot) \propto exp\left\{E_{\theta,z}\left[log(p(\phi|\beta)) + log(p(w|\phi,z))\right]\right\} \tag{6.7}$$

Using *log* on 6.5, and applying *log* to 6.3, and 6.4, we have:

$$log(q^*(z|\cdot)) \propto E_{\theta,\phi}\left[\sum_{d=1}^{D}\sum_{v=1}^{V}\sum_{k=1}^{K} n_{d,v} \times z_{d,v,k}\left(log(\theta_{d,k}) + log(\phi_{k,v})\right)\right]$$

$$\propto \sum_{d=1}^{D}\sum_{v=1}^{V}\sum_{k=1}^{K} n_{d,v} \times z_{d,v,k}\left(E(log(\theta_{d,k})) + E(log(\phi_{k,v}))\right)$$

Note that

$$x|p \sim Multinomial_K(p) \iff log\left(p(x|p)\right) = \sum_{k=1}^{K} x_k log(p_k),$$

and let's define $log(p_k) = E(log(\theta_{d,k}) + E(log(\phi_{k,v}))$, so $p_k = exp(E(log(\theta_{d,k})) + E(log(\phi_{k,v})))$.

Thus,

$$q^*(z|\cdot) \propto \prod_{d=1}^{D}\prod_{v=1}^{V}\prod_{k=1}^{K}\left[exp(E(log(\theta_{d,k})) + E(log(\phi_{k,v})))\right]^{n_{d,v} \times z_{d,v,k}}$$

That is,

$$z_{d,v}|w_d, \theta_d, \phi_k \sim Multinomial_K(p), \tag{6.8}$$

where $p = [p_1, p_2 \cdots, p_K]$; and by the Multinomial properties,

$$E(z_{d,v,k}) = p_k = exp(E(log(\theta_{d,k})) + E(log(\phi_{k,v}))) \tag{6.9}$$

Using Eq. 6.6, and taking the log of Eq. 6.2, and Eq. 6.3,

$$\begin{aligned} q^*(\theta|\cdot) &\propto exp\left\{E_z\left[\sum_d\sum_k(\alpha-1)log(\theta_{d,k}) + \sum_d\sum_k\sum_v n_{d,v}\times z_{d,v,k}log(\theta_{d,k})\right]\right\} \\ &= \prod_d^D\prod_{k=1}^K exp\left\{(\alpha + \sum_{v=1}^V n_{d,v}\times E(z_{d,v,k}) - 1)log(\theta_{d,k})\right\} \\ &= \prod_{d=1}^D\prod_{k=1}^K \theta_{d,k}^{\alpha+\sum_{v=1}^V n_{d,v}\times E(z_{d,v,k})-1} \end{aligned}$$

Thus, the approximate posterior distribution of the topics distribution in a document d is:

$$\theta_d|w_d, \alpha \sim Dirichlet_K(\tilde{\alpha}_d) \tag{6.10}$$

where $\tilde{\alpha}_d = \alpha + \sum_{v=1}^V n_{d,v}\times E(z_{d,v,\cdot})$. $\tilde{\alpha}_d$ is a K vector, and the dot $(\cdot)$ in $E(z_{d,v,\cdot})$ stands for K vector. The k^{th} element of the K vector $\tilde{\alpha}_d$ is $\tilde{\alpha}_{d,k} = \alpha_k + \sum_{v=1}^V n_{d,v}\times E(z_{d,v,k})$.

By the properties of the Dirichlet distribution, the expected value of $\theta_d|\tilde{\alpha}_d$ is given by:

$$E(\theta_d|\tilde{\alpha}_d) = \frac{\alpha + \sum_{v=1}^V n_{d,v}\times E(z_{d,v,\cdot})}{\sum_{k=1}^K[\alpha + \sum_{v=1}^V E(z_{d,v,k})]} \tag{6.11}$$

The numerical estimation of $E(\theta_d|\tilde{\alpha}_d)$ gives the estimates of the topics proportions within each document d, $(\hat{\theta}_d)$. It is worth noting that $E(z_{d,v,k})$ can be interpreted as the responsibility that topic k takes for explaining the observation of the word v in document d. Ignoring for a moment

the denominator of Eq. (6.11), $E(\theta_{d,k}|\tilde{\alpha}_{d,k})$ is similar to a regression equation where $n_{d,v}$ are the observed counts of words in document d, and $E(z_{d,v,k})$ are the parameter estimates (or weight) of the words. That illustrates that the importance of a topic in a document is due to the high presence of words $(n_{d,v})$ referring to that topic, and the weight of these words $(E(z_{d,v,k}))$.

Using Eq. 6.7, and taking the *log* of Eq. 6.1, and Eq. 6.4,

$$\begin{aligned} q^*(\phi) &\propto exp\left\{E_z\left[\sum_{k=1}^{K}\sum_{v=1}^{V}(\beta-1)log(\phi_{k,v}) + \sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{v=1}^{V} n_{d,v}\times z_{d,v,k} log(\phi_{k,v})\right]\right\} \\ &= \prod_{k=1}^{K}\prod_{v=1}^{V} exp\left\{(\beta + \sum_{d=1}^{D} n_{d,v}\times E(z_{d,v,k}) - 1)log(\phi_{k,v})\right\} \\ &= \prod_{k=1}^{K}\prod_{v=1}^{V} \phi_{k,v}^{\beta+\sum_{d=1}^{D} n_{d,v}\times E(z_{d,v,k})} \end{aligned}$$

Thus, the approximate posterior distribution of the words distribution in a topic $\hat{\phi}_k$ is:

$$\phi_k|w,\beta \sim Dirichlet_V(\tilde{\beta}_k) \tag{6.12}$$

where $\tilde{\beta}_k = \beta + \sum_{d=1}^{D} n_{d,v}\times E(z_{d,\cdot,k})$. $\tilde{\beta}_k$ is a V vector, and the dot $(\cdot)$ in $E(z_{d,\cdot,k})$ stands for V vector. The v^{th} element of the V vector $\tilde{\beta}_k$ is $\tilde{\beta}_{v,k} = \beta_v + \sum_{d=1}^{D} n_{d,v}E(z_{d,v,k})$.

And the expected value of $\phi_k|\tilde{\beta}_k$ is given by:

$$E(\phi_k|\tilde{\beta}_k) = \frac{\beta + \sum_{d=1}^{D} n_{d,v}\times E(z_{d,\cdot,k})}{\sum_{v=1}^{V}(\beta + \sum_{d=1}^{D} n_{d,v}\times E(z_{d,v,k}))} \tag{6.13}$$

The numerical estimation of $E(\phi_k|\tilde{\beta}_k)$ gives the estimates of the words relative importance for each topic k, (ϕ_k). Ignoring the denominator in the Eq. 6.13, $E(\phi_{k,v}|\tilde{\beta}_{k,v})$ is the weighted sum of the the frequencies of the word v in each of the documents $(n_{d,v})$, the weights being the responsibility topic k takes for explaining the observation of the word v in document d $(E(z_{d,v,k}))$.

Here, we have derived the posteriors expected values of the θs and ϕs using the words counts $n_{d,v}$,

which is slightly different from Blei et al. (2003). Posterior formulas similar to the current derived solution can be found in Murphy (2012, p.962).

In sum, the rows of $\phi_{K,V} = \left[E(\phi_k|\tilde{\beta}_k)\right]_{K\times V}$ are useful for interpreting (or identifying) the themes, which relative importance in each document are represented by the columns of $\theta_{D,K} = [E(\theta_d|\tilde{\alpha}_d)]_{D\times K}$.

Table 6: Comparing the parameter estimators of NMF, PLSA, and LDA

	Document topics distribution	**Topic words distribution**
NMF	$\hat{Z}_{d,k} = \sum_{v=1}^{V} W_{d,v} Q_{v,k}$	$\hat{B}_{k,v} = \sum_{d=1}^{D} W_{d,v} P_{k,d}$
PLSA	$p(z_k\|d_i) = \frac{\sum_{v=1}^{V} n(d_i,w_v) p(z_k\|d_i,w_v)}{\sum_{k=1}^{K}\sum_{v=1}^{V} n(d_i,w_v) p(z_k\|d_i,w_v)}$	$p(w_v\|z_k) = \frac{\sum_{d=1}^{D} n(d_i,w_v) p(z_k\|d_i,w_v)}{\sum_{v=1}^{V}\sum_{d=1}^{D} n(d_i,w_v) p(z_k\|d_i,w_v)}$
LDA	$E(\theta_d\|\tilde{\alpha}_d) = \frac{\alpha+\sum_{v=1}^{V} n_{d,v} E(z_{d,v,.})}{\sum_{k=1}^{K}[\alpha+\sum_{v=1}^{V} E(z_{d,v,k})]}$	$E(\phi_k\|\tilde{\beta}_k) = \frac{\beta+\sum_{d=1}^{D} n_{d,v} * E(z_{d,.,k})}{\sum_{v=1}^{V}(\beta+\sum_{d=1}^{D} n_{d,v} * E(z_{d,v,k}))}$

Table 6 compares the parameters estimators of NMF, PLSA, and LDA. To clarify notations, note that $W_{d,v} = n(d_i, w_v) = n_{d,v}$ is the count of the word w_v in document d_i. The difference between the three methods are mostly due to the differences in words counts weighting scheme. The hyperparameters, present in the LDA estimators, are also sources of differences.

7 Correlated Topic Model

LDA is what we would call the OLS of topic modeling. Current development of topic modeling generally consists of modifying the LDA assumptions. For example, Correlated Topic Modeling (CTM) modifies the prior assumption of the documents topic distribution θ. While LDA assumes that θ follows a Dirichlet prior distribution, CTM assumes that θ follows a normal (or Gaussian) prior distribution. Though the assumption of normal distribution is more realistic than the as-

sumption of Dirichlet distribution, the normal assumption renders the estimation of CTM more challenging. Since $z|\theta$ is multinomial, assuming Dirichlet distribution prior for θ is convenient because Dirichlet is a conjugate prior for the multinomial distribution. A drawback of the Dirichlet prior is that it assumes independence between the topics. However, this independence assumption is not realistic, since we expect topics within documents to be correlated. For instance it is likely that a topic about health is correlated with the topic about food, and the topic about exercise. CTM was proposed to allow correlation between topics within documents (Blei and Lafferty, 2007).

7.1 Model Specification

Figure 7.1: Path Diagram of the Correlated Topic Model

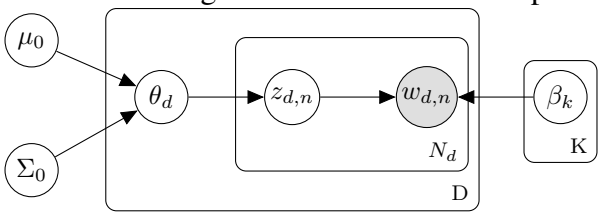

Figure 1 presents the path diagram of the correlated topic model. The diagram shows the conditional links between the variables. The arrows show the directions of the conditional links. For instance, θ_d is conditional on μ_0, and Σ_0; $z_{d,n}$ is conditioned on θ_d.

From the diagram, the joint distribution of θ_d, z_d, w_d, that is, the distribution of jointly observing θ_d, z_d, w_d is:

$$p(\theta_d, z_d, w_d|\mu_0, \Sigma_0, \beta) = p(\theta_d|\mu_0, \Sigma_0)p(z_d|\theta_d)p(w_d|\beta, z_d)$$

Note: only the vector of words w_d counts is observed. θ_d and z_d are unobserved (or hidden) variables. The goal is to infer the posterior distribution of θ_d and z_d conditional on the observed words w_d[1].

[1]Here, we derive the posterior distribution for the parameters of a single document.

$$p(\theta_d, z_d | w_d, \mu_0, \Sigma_0, \beta) = \frac{p(\theta_d, z_d, w_d | \mu_0, \Sigma_0, \beta)}{\int \sum_z [p(\theta_d, z_d, w_d | \mu_0, \Sigma_0, \beta)] d\theta} \tag{7.1}$$

Because of the complexity of the denominator of Eq. 7.1, this posterior distribution is intractable. Hence, the use of iterative methods to approximate the posterior distribution of θ, and z. Here, the Variational Inference (VI) method is used. Before, we derive the variational approximation of the posteriors, let's spell out the distribution of the random variables θ_d, z_d, and w_d.

The distributional assumptions are:

$$\theta_d | \mu_0, \Sigma_0 \sim Normal(\mu_0, \Sigma_0)$$

i.e.

$$p(\theta_d | \mu_0, \Sigma_0) = |2\pi\Sigma_0|^{-1/2} exp\{-\frac{1}{2}(\theta_d - \mu_0)^T \Sigma_0^{-1} (\theta_d - \mu_0)\} \tag{7.2}$$

$$z_{d,n} | \theta_d \sim Multinomial(1, \pi(\theta_d))$$

where

$$\pi(\theta_d) = \frac{exp(\theta_d)}{\sum_l^K exp(\theta_{d,l})}$$

and

$$\sum_{k=1}^{K} \pi(\theta_{d,k}) = 1$$

So,

$$p(z_{d,n} = k | \theta_d) = \pi(\theta_{d,k})$$

and

$$p(z_{d,n} | \theta_d) = \prod_{k=1}^{K} \left[\pi(\theta_{d,k})\right]^{z_{d,n,k}},$$

where $z_{d,n}$ is a K vector of zeros and 1 at the k^{th} position ($z_{d,n} = [0, 0, \ldots, 0, 1, \ldots]$). K is the number

of topics, arbitrarily chosen. n refers to the n^{th} word. Since document d has N_d independent words,

$$p(z_d|\theta_d) = \prod_{n=1}^{N_d} \left[\prod_{k=1}^{K} \left(\pi(\theta_{d,k})\right)^{z_{d,n,k}} \right] \tag{7.3}$$

$$w_{d,n}|\beta, z_{d,n} \sim Multinomial(\beta_{k=z_{d,n}})$$

i.e.

$$p(w_{d,n} = v|\beta_{k=z_{d,n,k}}) = \beta_{k,v}$$

and,

$$p(w_{d,n}|\beta, z_d) = \prod_{k=1}^{K} \left[\prod_{v=1}^{V} \beta_{k,v}^{w_{d,n,v}} \right]^{z_{d,n,k}}$$

where $w_{d,n}$ is a V vector of zeros and 1 at the v^{th} position ($w_{d,n} = [0,0,\ldots,0,1,\ldots]$). V is the number of unique words. Since document d has N_d independent words,

$$p(w_d|\beta, z_d) = \prod_{n=1}^{N_d} \left\{ \prod_{k=1}^{K} \left[\prod_{v=1}^{V} \beta_{k,v}^{w_{d,n,v}} \right]^{z_{d,n,k}} \right\} \tag{7.4}$$

Recall that the joint distribution of w_d, z_d and θ_d is given by:

$$p(\theta_d, z_d, w_d|\mu_0, \Sigma_0, \beta) = p(\theta_d|\mu_0, \Sigma_0)p(z_d|\theta_d)p(w_d|\beta, z_d)$$

and can be written explicitly by taking the product of Eq. 7.2, Eq. 7.3, and Eq. 7.4.

7.2 Variational Bayes

Eq. 7.1 reveals that solving for the posterior distribution of the latent variables is intractable due to the difficulty of computing the marginal distribution of the observed data w_d. MCMC and Variational Bayes (VB) methods can be used to approximate the posterior. VB is a faster algorithm, especially when the model becomes complex. Blei and Lafferty (2007) and Wang and Blei (2013)

apply VB to approximate the posterior distributions of θ and z.

By Bishop (2006, p.462), the variational approximation of the posterior distributions can be written as:

$$q^*(\theta_d|\cdot) \propto exp\left\{E_{q(z_d)}\left[log\left(p(z_d|\theta_d)p(\theta_d|\mu_0,\Sigma_0)\right)\right]\right\}$$

$$q^*(z_d|\cdot) \propto exp\left\{E_{q(\theta)}\left[log\left(p(z_d|\theta_d)p(w_d|\beta,z_d)\right)\right]\right\}$$

The dot in $|\cdot$ is a place holder for posterior parameters. VB relies on the hope that $exp\{\cdot\}$ (where $\cdot$ is the expression inside the exponential) will take the form of a known distribution; and the distribution $q^*(\cdot)$ is inferred to be that known distribution. The following development will make this idea clearer.

Deriving $q^*(\theta_d|\cdot)$

$$\begin{aligned} q^*(\theta_d|\cdot) &\propto exp\left\{E_{q(z_d)}\left[log\left(p(z_d|\theta_d)p(\theta_d|\mu_0,\Sigma_0)\right)\right]\right\} \\ &= exp\left\{E_{q(z_d)}\left[log\left(p(z_d|\theta_d)\right)\right] + log\left(p(\theta_d|\mu_0,\Sigma_0)\right)\right\} \end{aligned} \tag{7.5}$$

$$\begin{aligned} log\left(p(z_d|\theta_d)\right) &= \sum_{n=1}^{N_d}\sum_{k=1}^{K} z_{d,n,k} log\left(\pi(\theta_{d,k})\right) \\ &= \sum_{n=1}^{N_d}\sum_{k=1}^{K} z_{d,n,k}\left[\theta_{d,k} - log\left(\sum_{l=1}^{K} exp\left\{\theta_{d,k}\right\}\right)\right] \end{aligned} \tag{7.6}$$

$$log\left(p(\theta_d|\mu_0,\Sigma_0)\right) \propto -\frac{1}{2}(\theta_d-\mu_0)^T\Sigma^{-1}(\theta_d-\mu_0) \tag{7.7}$$

Thus, Eq.7.6 and Eq.7.7 into Eq.7.5 gives:

$$q^*(\theta_d|\cdot) \propto exp\left\{E_{q(z_d)}\left[\sum_{n=1}^{N_d}\sum_{k=1}^{K} z_{d,n,k}\left[\theta_{d,k} - log\left(\sum_{l=1}^{K} exp\left\{\theta_{d,k}\right\}\right)\right]\right] - \frac{1}{2}(\theta_d-\mu_0)^T\Sigma^{-1}(\theta_d-\mu_0)\right\}$$

Let $f(\theta_d)$ be the expression inside the exponential, that is,

$$\begin{aligned} f(\theta_d) &= \sum_{n=1}^{N_d}\sum_{k=1}^{K} E_{q(z_d)}(z_{d,n,k})\left[\theta_{d,k} - log\left(\sum_{l=1}^{K} exp\left\{\theta_{d,l}\right\}\right)\right] - \frac{1}{2}(\theta_d-\mu_0)^T\Sigma_0^{-1}(\theta_d-\mu_0) \\ &= \left[\theta_d - log(\sum_{l=1}^{K} exp\{\theta_{d,l}\})\right]^T E_{q(z_d)}\left(\sum_{n=1}^{N_d} z_{d,n.}\right) - \frac{1}{2}(\theta_d-\mu_0)^T\Sigma_0^{-1}(\theta_d-\mu_0) \qquad (7.8) \\ &= \left[\theta_d - log\left(\sum_{l=1}^{K} exp\left\{\theta_{d,l}\right\}\right)\right]^T E_{q(z_d)}\left(t(z_d)\right) - \frac{1}{2}(\theta_d-\mu_0)^T\Sigma_0^{-1}(\theta_d-\mu_0) \end{aligned}$$

where $t(z_d) = \sum_{n=1}^{N_d} z_{d,n,\cdot}$, and the dot $(\cdot)$ in $z_{d,n,\cdot}$ stands for K-vector.

Since $exp\{f(\theta_d)\}$ does not have the form of a distribution we can readily identify, we will approximate it by the normal distribution, using the Laplace approximation method.

The Laplace approximation uses the 2^{nd} order Taylor series expansion of $f(\theta_d)$.

By the 2^{nd} order Taylor expansion,

$$f(\theta_d) \approx f(\hat{\theta}_d) + \bigtriangledown f(\hat{\theta}_d)^T(\theta_d - \hat{\theta}_d) + \frac{1}{2}(\theta_d - \hat{\theta}_d)^T[\bigtriangledown^2 f(\hat{\theta}_d)](\theta_d - \hat{\theta}_d)$$

Where

$$\hat{\theta}_d = \arg\max_{\theta_d} f(\theta_d)$$

$\bigtriangledown f(\hat{\theta}_d)$ is the gradient of f, and $[\bigtriangledown^2 f(\hat{\theta}_d)]$ is the hessian of f, both evaluated at $\hat{\theta}_d$.

Note that from the first order condition (FOC), $\bigtriangledown f(\hat{\theta}_d) = 0$; therefore,

$$f(\theta_d) \approx f(\hat{\theta}_d) - \frac{1}{2}\left(\theta_d - \hat{\theta}_d\right)^T\left[-\bigtriangledown^2 f(\hat{\theta}_d)\right]\left(\theta_d - \hat{\theta}_d\right) \qquad (7.9)$$

$exp\{f(\theta_d)\}$ can thus be approximated as:

$$exp\{f(\theta_d)\} \propto C \times exp\left\{-\frac{1}{2}\left(\theta_d - \hat{\theta}_d\right)^T \left[-\bigtriangledown^2 f(\hat{\theta}_d)\right]\left(\theta_d - \hat{\theta}_d\right)\right\}$$

Where $C = exp\left\{f(\hat{\theta}_d)\right\}$. So,

$$q^*(\theta_d|\cdot) \propto exp\{f(\theta_d)\} = C \times exp\left\{-\frac{1}{2}\left(\theta_d - \hat{\theta}_d\right)^T \left[-\bigtriangledown^2 f(\hat{\theta}_d)\right]\left(\theta_d - \hat{\theta}_d\right)\right\}$$

And we can readily identify the kernel of the multivariate normal density with mean $\hat{\theta}_d$ and variance $\Sigma = [-\bigtriangledown^2 f(\hat{\theta}_d)]^{-1}$

It can be shown (see Appendix C) that:

$$\bigtriangledown f(\theta_d) = E_{q(z_d)}[t(z_d)] - \pi(\theta_d)\sum_{k=1}^{K}\left[E_{q(z_d)}(t(z_d))\right]_k - \Sigma_0^{-1}(\theta_d - \mu_0)$$

And

$$\bigtriangledown^2 f(\theta_d) = \left[-diag(\pi(\theta_d)) + \pi(\theta_d)\pi(\theta_d)^T\right]\sum_{k=1}^{K}\left[E_{q(z_d)}(t(z_d))\right] - \Sigma_0^{-1}$$

To summarize, the variational approximation of the posterior distribution of θ_d is:

$$q^*(\theta_d|\mu,\Sigma) = Normal\left(\mu = \hat{\theta}_d, \Sigma = \left[-\bigtriangledown^2 f(\hat{\theta}_d)\right]^{-1}\right)$$

Where

$$\hat{\theta}_d = \arg\max_{\theta_d} f(\theta_d) \tag{7.10}$$

There is not a closed form solution for Eq. 7.10; we can use numerical optimization method, such as gradient conjugate method, to approximate $\hat{\theta}_d$.

Deriving $q^*(z_d|\cdot)$

Again,

$$
\begin{aligned}
p(z_d|\theta_d) &= \prod_{n=1}^{N_d}\prod_{k=1}^{K}\pi(\theta_{d,k})^{z_{d,n,k}} = exp\left\{\sum_{n=1}^{N_d}\sum_{k=1}^{K} z_{d,n,k} log(\pi(\theta_{d,k}))\right\} \\
&= exp\left\{\sum_{n=1}^{N_d}\sum_{k=1}^{K} z_{d,n,k} log\left(\frac{exp(\theta_{d,k})}{\sum_{l=1}^{K} exp(\theta_{d,l})}\right)\right\} \\
&= exp\left\{\sum_{n=1}^{N_d}\sum_{k=1}^{K} z_{d,n,k}\left[\theta_{d,k} - log(\sum_{l=1}^{K} exp(\theta_{d,l})\right]\right\} \\
&= exp\left\{\sum_{k=1}^{K}\sum_{n=1}^{N_d} z_{d,n,k}\left[\theta_{d,k} - log(\sum_{l=1}^{K} exp(\theta_{d,l})\right]\right\} \\
&= exp\left\{\sum_{k=1}^{K} t(z_d)_k\left[\theta_{d,k} - log\left(\sum_{l=1}^{K} exp(\theta_{d,l})\right)\right]\right\} \\
&= exp\left\{\left[\theta_d - log\left(\sum_{l=1}^{K} exp(\theta_{d,l})\right)\right]^T t(z_d)\right\}
\end{aligned}
$$

So,

$$
\begin{aligned}
log\left(p(z_d|\theta_d)\right) &= \left[\theta_d - log\left(\sum_{l=1}^{K} exp(\theta_{d,l})\right)\right]^T t(z_d) \\
&= \eta(\theta_d)^T t(z_d)
\end{aligned}
$$

Where $\eta(\theta_d) = \left[\theta_d - log\left(\sum_{l=1}^{K} exp(\theta_{d,l})\right)\right]$

Thus,

$$
E_{q(\theta)}\left\{log\left(p(z_d|\theta_d)\right)\right\} = \left[E_{q(\theta)}\left\{\eta(\theta_d)\right\}\right]^T t(z_d) \tag{7.11}
$$

$$
\begin{aligned}
p(w_d|\beta, z_d) &= \prod_{n=1}^{N_d}\left\{\prod_{k=1}^{K}\left(\prod_{v=1}^{V}\beta_{k,v}^{w_{d,n,v}}\right)^{z_{d,n,k}}\right\} \\
&= exp\left\{\sum_{n=1}^{N_d}\sum_{k=1}^{K} z_{d,n,k}\sum_{v=1}^{V} w_{d,n,v} log(\beta_{k,v})\right\}
\end{aligned}
$$

Note that for each n, $w_{d,n}$ is a $V \times 1$ vector of zeros and 1 at the v^{th} position. So $\sum_{v=1}^{V} w_{d,n,v} log(\beta_{k,v}) = log(\beta_{k,v})$. Thus,

$$\begin{aligned} p(w_d|\beta, z_d) &= exp\left\{\sum_{k=1}^{K}\left[\sum_{n=1}^{N_d} z_{d,n,k} log(\beta_{k,v})\right]\right\} \\ &= exp\left\{\sum_{k=1}^{K}\left[t(z_d)_k log(\beta_{k,v})\right]\right\} \\ &= exp\left\{t(z_d)^T log(\beta_{\cdot,v})\right\} \\ &= exp\left\{log(\beta_{\cdot,v})^T t(z_d)\right\} \\ &= exp\left\{t(w_d)^T t(z_d)\right\} \end{aligned}$$

where $t(z_d) = \sum_{n=1}^{N_d} z_{d,n,\cdot}$, and $t(w_d) = log(\beta_{\cdot,v})$ is a $K \times 1$ vector.

So,

$$log\left(p(w_d|\beta, z_d)\right) = t(w_d)^T t(z_d) \tag{7.12}$$

And

$$q^*(z_d|\cdot) \propto exp\left\{\left[E_{q(\theta)}\left(\eta(\theta_d)\right) + t(w_d)\right]^T t(z_d)\right\}$$

$$\begin{aligned} E_{q(\theta)}\left(\eta(\theta_d)\right) &= E_{q(\theta)}\left[\theta_d - log\left(\sum_{l=1}^{K} exp(\theta_{d,l})\right)\right] \\ &= \mu_d - log\left(\sum_{l=1}^{K} exp(\mu_{d,l})\right) \end{aligned}$$

(Note: $\mu_d = \hat{\theta}_d$)

By Wang and Blei (2013), (equation 17),

$$q^*(z_d|\phi_d) \sim Multinomial_K(\phi_d) \tag{7.13}$$

where ϕ_d is a $K \times V$ matrix with columns

$$\phi_{d,\cdot,v} = exp\left\{\mu_d - log\left(\sum_{l=1}^{K} exp(\mu_{d,l})\right) + log\left(\beta_{\cdot,v}\right)\right\}$$

To see why this is the case, observe that if $\underline{X} \sim Multinomial(\underline{p})$, then $p(\underline{X}|\underline{p}) \propto \prod_{k=1}^{K} p_i^{X_i} = exp\left\{\sum_{k=1}^{K} X_i log(p_i)\right\} = exp\left\{log(\underline{p})^T \underline{X}\right\}$

Thus, the elements of the ϕ_d matrix are:

$$\begin{aligned}\phi_{d,k,v} &= exp\left\{\mu_{d,k} - log\left(\sum_{l=1}^{K} exp(\mu_{d,l})\right) + log\left(\beta_{k,v}\right)\right\} \\ &= exp\left\{\mu_{d,k} - log\left(\sum_{l=1}^{K} exp(\mu_{d,l})\right)\right\}\beta_{k,v}\end{aligned}$$

To summarize, for a given word $w_{d,n} = v$,

$$q^*(z_{d,v,\cdot}) = Multinomial_K(\phi_{d,\cdot,v})$$

where $\phi_{d,\cdot,v} = exp\left\{\mu_d - log\left(\sum_{l=1}^{K} exp(\mu_{d,l})\right)\right\} \odot \beta_{\cdot,v}$ ($\odot$ indicates element wise vector multiplication).

Put differently, $\phi_{d,k,v} = exp\left\{\mu_{d,k} - log\left(\sum_{l=1}^{K} exp(\mu_{d,l})\right)\right\} \times \beta_{\cdot,v}$

$$t(z_d) = \sum_{n=1}^{N_d} z_{d,n,\cdot}$$

is a $K \times 1$ vector, and

$$E_{q(z_d)}(t(z_d)) = E_{q(z_d)}\left(\sum_{n=1}^{N_d} z_{d,n,\cdot}\right)$$
$$= \phi_d n_d^T$$

by the Multinomial properties (if $\underline{X} \sim Multinomial(\underline{p})$, then $E(X_i) = np_i$, n being the total number of observations).

We have presented topic modeling in evolutionary perspective, by linking LDA to NMF and PCA, which are easier to grasp, intuitively. Current topic modeling algorithms deviate from LDA by modifying the LDA assumptions. For example, CTM replaces the Dirichlet assumption of LDA with a logit normal distribution.

References

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., and Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, 52(1):155–173.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Eldén, L. (2007). *Matrix methods in data mining and pattern recognition*, volume 4. SIAM.

Gaussier, E. and Goutte, C. (2005). Relation between plsa and nmf and implications. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 601–602. ACM.

Gentle, J. E. (2017). *Matrix Algebra: Theory, Computations, and Applications in Statistics*. Springer Publishing Company, Incorporated, 2nd edition.

Gillis, N. (2017). Introduction to nonnegative matrix factorization. *arXiv preprint arXiv:1703.00663*.

Girolami, M. and Kabán, A. (2003). On an equivalence between plsi and lda. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 433–434. ACM.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 289–296, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Hubert, L., Meulman, J., and Heiser, W. (2000). Two purposes for matrix factorization: A historical appraisal. *SIAM review*, 42(1):68–82.

Landauer, T. K. and Dumais, S. T. (1997a). The latent semantic analysis theory of acquisition. *Induction and Representation of Knowledge1997*.

Landauer, T. K. and Dumais, S. T. (1997b). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Lu, Y., Mei, Q., and Zhai, C. (2011). Investigating task performance of probabilistic topic models: an empirical study of plsa and lda. *Information Retrieval*, 14(2):178–203.

Mimno, D. and McCallum, A. (2012). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278*.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Shlens, J. (2014). A tutorial on principal component analysis. *CoRR*, abs/1404.1100.

Taddy, M. (2012). On estimation and selection for topic models. In Lawrence, N. D. and Girolami, M. A., editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, volume 22, pages 1184–1193.

Thurstone, L. L. (1935). The vectors of mind: Multiple-factor analysis for the isolation of primary traits.

Wang, C. and Blei, D. M. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(Apr):1005–1031.

A Matrix Factorization Algorithm

A.1 Unconstrained Matrix Factorization

The following R code implements the matrix factorization algorithm on the example data provided.

```
W <- matrix(c(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,
                                2, 6, 1, 4, 5), byrow = TRUE, nrow = 6)
W <- data.frame(W)
names(W) = c("college", "education", "family", "health", "medicaid")
row.names(W) = paste0("document.", 1:6)
W = as.matrix(W)


set.seed(3)
Z_init <- abs(round(rnorm(n = 6*2, mean = 0, sd = 2),0))
Z_init <- matrix(Z_init, nrow = 6)
Z <- Z_init
dist_ww <- 1e3
max_iter <- 1000
iter <- 0
while(iter <= max_iter && dist_ww >= 1e-6) {
        iter <- iter + 1
        ZZ_inv <- solve(t(Z)%*%Z)
        B <- ZZ_inv%*%t(Z)%*%W
```

```
        BB_inv <- solve(B%*%t(B))
        Z <- W%*%t(B)%*%BB_inv
        W_hat <- Z%*%B
        dist_ww <- sqrt(sum(W-W_hat)^2)
}

Z <- data.frame(round(Z, 2))
names(Z) <- c("Topic.1", "Topic.2")
B <- data.frame(round(B, 2), row.names = c("Topic.1", "Topic.2"))
```

Below is the table of the least squares estimate of B:

```
B
```

```
##         college education family health medicaid
## Topic.1    1.18      1.96  -0.02    0.6     0.58
## Topic.2    0.50      0.85   1.11    2.5     3.60
```

Observe that row 1 of B has high values in columns 1 and 2 compared to columns 3, 4, and 5; and row 2 has higher values for columns 4 and 5 compared to columns 1, 2, and 3. It is reasonable to infer that ***Topic.1*** (or component 1) is about education, and ***Topic.2*** is about health.

Below is the table of the least squares estimate of Z:

```
Z
```

```
##            Topic.1 Topic.2
## document.1    3.13    0.05
## document.2   -1.55    3.58
## document.3    4.31    0.97
```

```
## document.4    0.41    2.71
## document.5   -1.16    2.68
## document.6    2.26    1.03
```

Observe that ***Topic.1*** (or the scores of component 1) has big absolute values in documents 1, 3, and 6. Likewise, ***Topic.2*** has big values in documents 2, 4, and 5. Hence, we can infer that documents 1, 4, and 6 are mostly about education; and documents 2, 4, and 5 are mostly about health.

We can use a scatterplot to explore the original five dimensional W data in a two dimensional Z data as follow:

```
plot(x = Z$Topic.1, y = Z$Topic.2, cex = 3, xlab = "Topic.1", ylab = "Topic.2",
                        ylim = c(-0.25, 3.75), xlim = c(-1.75, 4.5))
text(x = Z$Topic.1, y = Z$Topic.2, labels= 1:6, cex= 1)
```

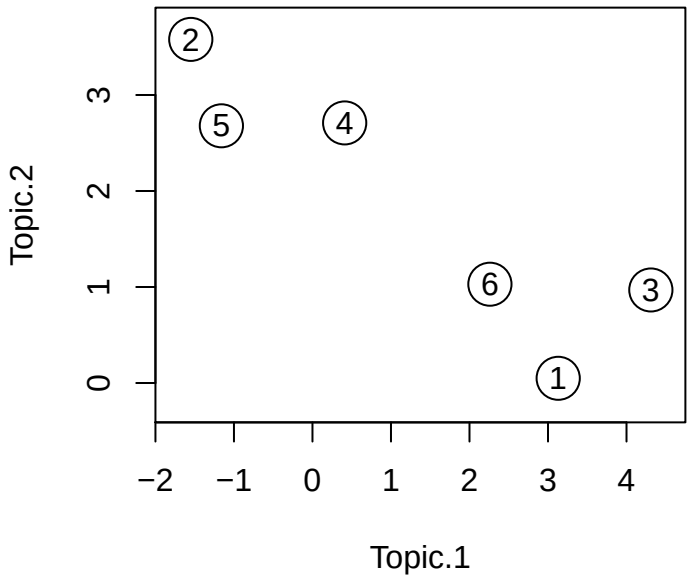

Figure A.1: Scatterplot of the two dimensional Z variables obtained from MF

Fig. 2.1 is similar to Fig. A.1 even though the Zs appear different. In fact, there is not a unique solution. Each solution is a local minimum, and is dependent on the initial value chosen. This non uniqueness of the solution poses some challenges for inferential analysis, just like Factor Analysis. For completeness, observe that $Z \times B$ approximates W well.

```
as.matrix(Z)%*%as.matrix(B)
```

```
##             college education  family health medicaid
## document.1  3.7184    6.1773 -0.0071   2.00     2.00
## document.2 -0.0390    0.0050  4.0048   8.02    11.99
## document.3  5.5708    9.2721  0.9905   5.01     5.99
## document.4  1.8388    3.1071  2.9999   7.02     9.99
## document.5 -0.0288    0.0044  2.9980   6.00     8.98
## document.6  3.1818    5.3051  1.0981   3.93     5.02
```

A.2 Non Negative Matrix Factorization Algorithm

The above R code can be modified to impose the non negative constraint, as shown below:

```
W <- matrix(c(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,
                                2, 6, 1, 4, 5), byrow = TRUE, nrow = 6)
W <- data.frame(W)
names(W) = c("college", "education", "family", "health", "medicaid")
row.names(W) = paste0("document.", 1:6)
```

```r
W = as.matrix(W)


set.seed(3)
Z_init <- abs(round(rnorm(n = 6*2, mean = 0, sd = 2),0))
Z_init <- matrix(Z_init, nrow = 6)
Z <- Z_init
dist_ww <- 1e3
max_iter <- 1000
iter <- 0
while(iter <= max_iter && dist_ww >= 1e-6) {
        iter <- iter + 1
        ZZ_inv <- solve(t(Z)%*%Z)
        B <- ZZ_inv%*%t(Z)%*%W
        B[B<0] <- 0 # impose non negative constraint
        BB_inv <- solve(B%*%t(B))
        Z <- W%*%t(B)%*%BB_inv
        Z[Z<0] <- 0 # impose non negative constraint
        W_hat <- Z%*%B
        dist_ww <- sqrt(sum(W-W_hat)^2)
}


Z <- data.frame(round(Z, 2))
names(Z) <- c("Topic.1", "Topic.2")
B <- data.frame(round(B, 2), row.names = c("Topic.1", "Topic.2"))
Z


##            Topic.1 Topic.2
```

```
## document.1    1.60    0.00
## document.2    0.00    5.83
## document.3    2.40    1.45
## document.4    0.79    4.37
## document.5    0.00    4.38
## document.6    1.37    1.60
```

```
B
```

```
##          college education family health medicaid
## Topic.1    2.33      3.85   0.00   1.26     1.25
## Topic.2    0.00      0.01   0.69   1.37     2.06
```

```
as.matrix(Z)%*%as.matrix(B)
```

```
##            college education family health medicaid
## document.1    3.73    6.1600   0.00   2.02     2.00
## document.2    0.00    0.0583   4.02   7.99    12.01
## document.3    5.59    9.2545   1.00   5.01     5.99
## document.4    1.84    3.0852   3.02   6.98     9.99
## document.5    0.00    0.0438   3.02   6.00     9.02
## document.6    3.19    5.2905   1.10   3.92     5.01
```

We observe that the matrices Z and B have non negative elements, as desired. The Z variables are plotted in Fig. A.2. The figure is very similar to Fig. A.1.

```
plot(x = Z$Topic.1, y = Z$Topic.2, cex = 3, xlab = "Topic.1", ylab = "Topic.2",
                        ylim = c(-0.25, 6.5), xlim = c(-0.25, 2.75))
text(x = Z$Topic.1, y = Z$Topic.2, labels= 1:6, cex= 1)
```

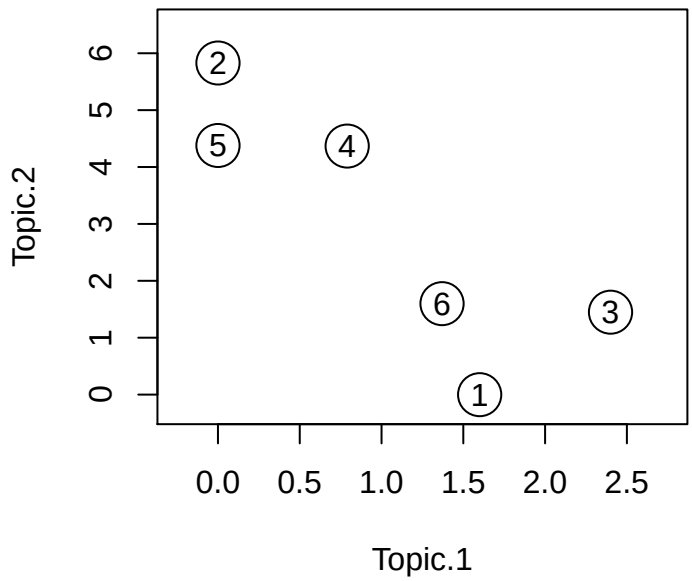

Figure A.2: Scatterplot of the two dimensional Z variables obtained from NMF

B Deriving the EM algorithm for PLSA

The Jensen Inequality states that if f is concave, then

$$E(f(x)) \leq f(E(x)) \tag{B.1}$$

Also, $\sum_{z \in Z} p(w_v|z)p(z|d_i)$ can be re-written as an expectation over some arbitrary random variable with probability distribution $q(z)$, as follows:

$$\sum_{z\in Z} p(w_v|z)p(z|d_i) = \sum_{z\in Z} q(z)p(w_v|z)p(z|d_i)\frac{1}{q(z)}$$
$$= E_{q(z)}\left(\frac{p(w_v|z)p(z|d_i)}{q(z)}\right)$$

Thus,

$$log\left(\sum_{z\in Z} p(w_v|z)p(z|d_i)\right) = log\left[E_{q(z)}\left(\frac{p(w_v|z)p(z|d_i)}{q(z)}\right)\right]$$

By Jensen inequality B.1, and by the fact that the log function is concave, we can write:

$$\begin{aligned} log\left[E_{q(z)}\left(\frac{p(w_v|z)p(z|d_i)}{q(z)}\right)\right] &\geq E_{q(z)}\left[log\left(\frac{p(w_v|z)p(z|d_i)}{q(z)}\right)\right] \\ &= E_{q(z)}\left[log\left(p(w_v|z)p(z|d_i)\right) - log\left(q(z)\right)\right] \\ &= E_{q(z)}\left[log[p(w_v|z)p(z|d_i)]\right] - E_{q(z)}\left[log\left(q(z)\right)\right] \\ &\geq E_{q(z)}\left[log\left(p(w_v|z)p(z|d_i)\right)\right] \end{aligned} \tag{B.2}$$

By eq. 5.5 and eq. B.2,

$$\begin{aligned} \mathscr{L}(\theta|W) &\geq \sum_{d=1}^{D}\sum_{v=1}^{V} n(d_i,w_v)E_{q(z)}\left[log\left(p(w_v|z)p(z|d_i)\right)\right] \\ &= \sum_{d=1}^{D}\sum_{v=1}^{V} n(d_i,w_v)\sum_{z\in Z} q(z)log\left(p(w_v|z)p(z|d_i)\right) \end{aligned} \tag{B.3}$$

Define $q(z_k) = p(z_k|w_v,d_i)$; then inequality B.3 becomes:

$$\mathscr{L}(\theta|W) \geq \sum_{d=1}^{D}\sum_{v=1}^{V} n(w_v,d_i)\sum_{k=1}^{K} p(z_k|w_v,d_i)log\left(p(w_v|z_k)p(z_k|d_i)\right) \tag{B.4}$$

Call the Right Hand Side (RHS) of inequality B.4 $Q(\theta)$. $Q(\theta)$ is a lower bound for $\mathscr{L}(\theta|W)$. The

θ that maximizes $Q(\theta)$ maximizes $\mathscr{L}(\theta|W)$ (Bishop, 2006, Chap.9).

$p(z_k|d_i)$ and $p(w_v|z_k)$ must satisfy the sum to 1 condition; i.e.

$$\sum_{k=1}^{K} p(z_k|d_i) = 1 \tag{B.5}$$

And

$$\sum_{v=1}^{V} p(w_v|z_k) = 1 \tag{B.6}$$

The EM goal is to find $p(z_k|d_i)$ and $p(w_v|z_k)$ to maximize $Q(\theta)$ subject to $\sum_{k=1}^{K} p(z_k|d_i) = 1$, and $\sum_{v=1}^{V} p(w_v|z_k) = 1$

We can now set the Lagrange optimization equation as:

$$\begin{aligned}\mathscr{L} = &\sum_{d=1}^{D}\sum_{v=1}^{V} n(d_i,w_v)\sum_{k=1}^{K} p(z_k|w_v,d_i)log\left(p(w_v|z_k)p(z_k|d_i)\right) \\ &+\sum_{k=1}^{K}\lambda_k\left(1-\sum_{v=1}^{V} p(w_v|z_k)\right)+\sum_{d=1}^{D}\tau_d\left(1-\sum_{k=1}^{K} p(z_k|d_i)\right)\end{aligned}$$

By the first order condition (FOC)

$$\frac{\partial\mathscr{L}}{\partial p(w_v|z_k)} = 0 \Leftrightarrow \sum_{d=1}^{D} n(d_i,w_v)p(z_k|d_i,w_v) = \lambda_k p(w_v|z_k) \tag{B.7}$$

Likewise,

$$\frac{\partial\mathscr{L}}{\partial p(z_k|d_i)} = 0 \Leftrightarrow \sum_{v=1}^{V} n(d_i,w_v)p(z_k|d_i,w_v) = \tau_d p(z_k|d_i) \tag{B.8}$$

By constraint B.5 and FOC B.7,

$$\sum_{v=1}^{V}\sum_{d=1}^{D} n(w_v,d_i)p(z_k|d_i,w_v) = \lambda_k \tag{B.9}$$

Likewise, by constraint B.6 and FOC B.8

$$\sum_{k=1}^{K}\sum_{v=1}^{V} n(w_v,d_i)p(z_k|d_i,w_v) = \tau_d \tag{B.10}$$

Substituting Eq. B.9 into B.7 yields:

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

$$\Longleftrightarrow$$

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

Likewise, substituting Eq. (B.10) into Eq. (B.8) yields:

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

$$\Longleftrightarrow$$

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

By Bayes rule,

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

To summarize, the EM algorithm for the PLSA consists of providing initial values for $p(w_v|z_k)$ and $p(z_k|d_i)$, then computing $p(z_k|d_i,w_v)$ in the E-Step as follows:

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

Once, $p(z_k|w_v,d_i)$ are computed, new values for $p(w_v|z_k)$ and $p(z_k|d_i)$ are computed in the M-Step

using:

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

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

To relate the PLSA estimates to the NMF estimate, the $[p(w_v|z_k)]_{K\times V}$ matrix is equivalent to the B matrix in NMF, and the $[p(z_k|d_i)]_{D\times K}$ matrix is equivalent to its Z.

C Deriving the first and second derivatives of $f(\theta_d)$

We have

$$\pi(\theta_d) = \frac{exp\{\theta_d\}}{\sum_{l=1}^{K} exp\{\theta_{d,l}\}}$$

and define

$$\eta(\theta_d) = log\,(\pi(\theta_d)) = \theta_d - log(\sum_{l=1}^{K} exp\{\theta_{d,l}\})$$

or

$$\pi(\theta_d) = exp\{\eta(\theta_d)\}$$

Define

$$A = \eta(\theta_d)^T \hat{t}(z_d)$$

where $\hat{t}(z_d) = E_{q(z_d)}(t(z_d))$

We need to use the property of vector-by-vector derivatives, which says that the derivative of a vector function $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{bmatrix}$ with respect to an input vector $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$ is given by:

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_1} \\ \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_1} \end{bmatrix} \tag{C.1}$$

Using Eq. C.1, it is easy to show that

$$\bigtriangledown A = \left[\hat{t}(z_d) - \pi(\theta_d) \sum_{k=1}^{K} \hat{t}(z_d)_k \right]$$

$$\frac{\partial \left[\frac{1}{2} (\theta_d - \mu_0)^T \Sigma_0^{-1} (\theta_d - \mu_0) \right]}{\partial \theta_d} = \Sigma_0^{-1} (\theta_d - \mu_0)$$

So,

$$\bigtriangledown f(\theta_d) = E_{q(z_d)} [t(z_d)] - \pi(\theta_d) \sum_{k=1}^{K} \left[E_{q(z_d)} (t(z_d)) \right]_k - \Sigma_0^{-1} (\theta_d - \mu_0)$$

We derive $\bigtriangledown f(\theta_d)$ with respect to θ_d to get the hessian matrix.

Again,

$$\pi(\theta_d) = exp\{\eta(\theta_d)\} = \frac{exp\{\theta_d\}}{\sum_{l=1}^{K} exp\{\theta_{d,l}\}}$$

$$\begin{aligned} \frac{\partial \pi(\theta_{d,i})}{\partial \theta_{d,i}} &= \frac{exp\{\theta_{d,i}\} \sum_{l=1}^{K} exp\{\theta_{d,l}\} - exp\{\theta_{d,i}\} exp\{\theta_{d,i}\}}{[\sum_{l=1}^{K} exp\{\theta_{d,l}]^2} \\ &= \pi(\theta_{d,i}) - \pi(\theta_{d,i}) \pi(\theta_{d,i}) \\ &= \pi(\theta_{d,i})[1 - \pi(\theta_{d,i})] \end{aligned}$$

$$\frac{\partial \pi(\theta_{d,i})}{\partial \theta_{d,j}} = -\frac{exp\{\theta_{d,i}\}exp\{\theta_{d,j}\}}{[\sum_{l=1}^{K} exp\{\theta_{d,l}]^2}$$
$$= -\pi(\theta_{d,i})\pi(\theta_{d,j})$$

So

$$\frac{\partial \pi(\theta_{d,i})}{\partial \theta_{d,j}} = \pi(\theta_{d,i})[1_{[i=j]} - \pi(\theta_{d,j}]$$

And, in matrix format,

$$\frac{\partial \pi(\theta_d)}{\partial \theta_d} = -diag(\pi(\theta_d)) + \pi(\theta_d)\pi(\theta_d)^T$$

Therefore,

$$\bigtriangledown^2 f(\theta_d) = \left[-diag\left(\pi(\theta_d)\right) + \pi(\theta_d)\pi(\theta_d)^T\right] \sum_{k=1}^{K} \left[E_{q(z_d)}(t(z_d))\right]_k - \Sigma_0^{-1}$$