SIMPLEX DECOMPOSITIONS USING SVD AND PLSA

Madhusudana Shashanka and Michael Giering
United Technologies Research Center, East Hartford, CT 06108, U.S.A.

Keywords: Matrix factorization, Probabilistic Latent Semantic Analysis (PLSA), Nonnegative Matrix Factorization (NMF), Singular Values Decomposition (SVD), Principal Components Analysis (PCA).

Abstract: Probabilistic Latent Semantic Analysis (PLSA) is a popular technique to analyze non-negative data where multinomial distributions underlying every data vector are expressed as linear combinations of a set of *basis distributions*. These learned basis distributions that characterize the dataset lie on the standard simplex and themselves represent corners of a simplex within which all data approximations lie. In this paper, we describe a novel method to extend the PLSA decomposition where the *bases* are not constrained to lie on the standard simplex and thus are better able to characterize the data. The locations of PLSA basis distributions on the standard simplex depend on how the dataset is aligned with respect to the standard simplex. If the directions of maximum variance of the dataset are orthogonal to the standard simplex, then the PLSA bases will give a poor representation of the dataset. Our approach overcomes this drawback by utilizing Singular Values Decomposition (SVD) to identify the directions of maximum variance, and transforming the dataset to align these directions parallel to the standard simplex before performing PLSA. The learned PLSA features are then transformed back into the data space. The effectiveness of the proposed approach is demonstrated with experiments on synthetic data.

1 INTRODUCTION

The need for analyzing non-negative data arises in several applications such as computer vision, semantic analysis and gene expression analysis among others. Nonnegative Matrix Factorization (NMF) (Lee and Seung, 1999; Lee and Seung, 2001) was specifically proposed to analyze such data where every data vector is expressed as a linear combination of a set of characteristic *basis vectors*. The weights with which these vectors combine differ from data point to data point. All entries of the basis vectors and the weights are constrained to be nonnegative. The nonnegativity constraint produces basis vectors that can only combine additively without any cross-cancellations and thus can be intuitively thought of as *building blocks* of the dataset. Given these desirable properties, the technique has found wide use across different applications. However, one of the main drawbacks of NMF is that the *energies* of data vectors is split between the basis vectors and mixture weights during decomposition. In other words, the basis vectors may lie in an entirely different part of the data space making any geometric interpretation meaningless.

Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 2001) is a related method with probabilistic foundations which was proposed around the same time in the context of semantic analysis of document corpora. A corpus of documents is represented as a matrix where each column vector corresponds to a document and each row corresponds to a word in the vocabulary and the entry corresponds to the numer of times the word appeared in the document. PLSA decomposes this matrix as a linear combination of a set of multinomial distributions over the words called *topics* where the weight vectors are multinomial distributions as well. Non-negativity constraint is imposed implicitly because the extracted topics or *basis distributions* and weights represent probabilities. It has been shown that the underlying computations in NMF and PLSA are identical (Gaussier and Goutte, 2005; Shashanka et al., 2008). However, unlike NMF where there are no additional constraints beyond nonegativity, PLSA bases and weights being multinomial distriutions also have the contraint that the entries sum to 1. Since the weights sum to 1, the PLSA approximations of the data can be thought of as lying within a simplex defined by the basis distriutions. (Shashanka, 2009) formalizes this geometric intuition as *Simplex Decompositions* where the model

Shashanka M. and Giering M. (2012).
SIMPLEX DECOMPOSITIONS USING SVD AND PLSA.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 248-252
DOI: 10.5220/0003797802480252

extracts basis vectors that combine additively and correspond to the corners of a simplex surrounding the modeled data. PLSA and its extensions such as Latent Dirichlet Allocation (Blei et al., 2003) and Correlated Topic Models (Blei and Lafferty, 2006) are specific examples of Simplex Decompositions.

Since PLSA (and other PLSA extensions) does not decompose the data-vectors themselves but the underlying multinomial distributions (i.e. the data vectors normalized to sum to unity), the extracted basis vectors don't lie in the data space but lie on the standard simplex. This can be a drawback depending on the dataset under consideration and may pose a particular poroblem if the data is aligned such that most of the variability and structure characterizing the dataset lies in directions orthogonal to the standard simplex. In such cases, the projections of the data vectors onto the simplex (which is what is decomposed by PLSA) carry very little information about the shape of the data distribution and thus the obtained PLSA bases are much less informative.

In this paper, we propose an approach to get around this drawback of PLSA. We first use Singular Values Decomposition (SVD) to identify the directions of the most variability in the dataset and then transform the dataset so that these vectors are parallel to the standard simplex. We perform PLSA on the transformed data and obtain PLSA basis vectors in the transformed space. Since the transformation is affine and invertible, we apply the inverse transformation on the basis vectors to obtain basis vectors the characterize the data in the original data space. These basis vectors no longer are constrained to live on the standard simplex but lie within the data space and correspond to corners of a simplex that surrounds all the data points.

The paper is organized as follows. In Section 2, we provide the necessary background by describing the PLSA algorithm and geometry. Section 3 describes our proposed approach and constitutes the bulk of the paper. We illustrate the applicability of the method by applying the proposed technique on synthetic data. We also provide a short discussion of the algorithm and its applicability for semi-nonnegative factorizations. We conclude the paper in Section 4 with a brief summary and avenues for future work.

2 BACKGROUND

Consider an $M \times N$ non-negative data matrix $\mathbf{V}$ where each column $\mathbf{v}_n$ represents the n-th data vector and v_{mn} represents the (mn)-th element. Let $\bar{\mathbf{v}}_n$ represent the normalized vector $\mathbf{v}_n$ and $\bar{\mathbf{V}}$ is the matrix $\mathbf{V}$ with all columns normalized.

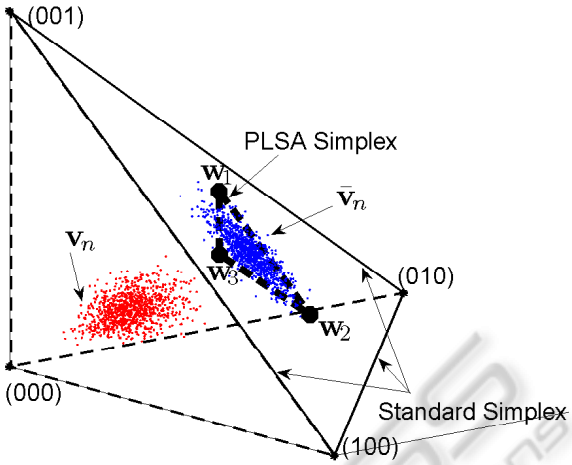

Figure 1: Illustration of Probabilistic Latent Semantic Analysis. The data matrix $\mathbf{V}$ with 1000 3-dimensional vectors $\mathbf{v}_n$ is shown as red points and the normalized data $\bar{\mathbf{V}}$ is shown as blue points on the Standard simplex. PLSA was performed on $\mathbf{V}$ and the three extracted basis distributions shown by $\mathbf{w}_1$, $\mathbf{w}_2$ and $\mathbf{w}_3$ are points on the Standard simplex that form the corners of the PLSA simplex around normalized data points $\bar{\mathbf{v}}_n$ shown in blue.

PLSA characterizes the bidimensional distribution $P(m,n)$ underlying $\mathbf{V}$ as

$$P(m,n) = P(n)P(m|n) = P(n)\sum_z P(m|z)P(z|n), \quad (1)$$

where z is a latent variable. PLSA represents $\bar{\mathbf{v}}_n$ as data distributions $P(m|n)$ which in turn is expressed as a linear combination of *basis distributions* $P(m|z)$. These basis distributions combine with different proportions given by $P(z|n)$ to form data distributions.

PLSA parameters $P(m|z)$ and $P(z|n)$ can be estimated through iterations of the following equations derived using the EM algorithm,

$$\begin{aligned} P(z|m,n) &= \frac{P(m|z)P(z|n)}{\sum_z P(m|z)P(z|n)}, \\ P(m|z) &= \frac{\sum_n v_{mn}P(z|m,n)}{\sum_m \sum_n v_{mn}P(z|m,n)}, \quad \text{and} \\ P(z|n) &= \frac{\sum_m v_{mn}P(z|m,n)}{\sum_m v_{mn}}. \end{aligned}$$

EM algorithm guarantees that the above updates converge to a local optimum.

PLSA can be written as a matrix factorization

$$\bar{\mathbf{V}}_{M\times N} \approx \mathbf{W}_{M\times Z}\mathbf{H}_{Z\times N} = \mathbf{P}_{M\times N}, \quad (2)$$

where $\mathbf{W}$ is the matrix of basis distributions $P(m|z)$ with column $\mathbf{w}_z$ corresponding to the z-th basis distribution, $\mathbf{H}$ is the mixture weight distriution matrix of entries $P(z|n)$ with column $\mathbf{h}_n$ corresponding to the

n-th data vector, and $\mathbf{P}$ is the matrix of model approximations $P(m|n)$ with column $\mathbf{p}_n$ corresponding to the n-th data vector. See Figure 1 for an illustration of PLSA.

3 ALGORITHM

The previous section described PLSA algorithm and illustrated the geometry of the technique. This section presents our proposed approach. We first briefly present the motivation for our algorithm and then describe the details of the algorithm. We illustrate the algorithm by applying it on a synthetic dataset.

3.1 Motivation

As illustrated in Figure 1, the basis distributions obtained by applying PLSA on a dataset lie on the Standard simplex. The basis distributions form the corners of a *PLSA Simplex* containing not the original datapoints but the normalized datapoints instead.

Our goal is to extend the technique so that the basis vectors form a simplex around the original datapoints. In other words, we would like to remove the constraint that the basis vectors form multinomial distributions and thus they don't have to lie on the standard simplex. However, since we need the basis vectors to still form a simplex around the data approximations, the mixture weights with which they combine are still constrained to be multinomial distributions.

The necessity of such an approach becomes apparent when one considers the implication of normalization of datapoints that PLSA implicitly does. The normalization skews the relative geometry of datapoints. In certain cases, the normalization can hide the real shape of the distriution of datapoints as illustrated in Figure 2.

3.2 Problem Formulation

Given the data matrix $\mathbf{V}$, we would like to find a matrix decomposition similar to equation 2 of the form

$$\mathbf{V}_{M\times N} \approx \mathcal{W}_{M\times Z}\mathcal{H}_{Z\times N} = \mathcal{P}_{M\times N} \qquad (3)$$

where Z is the dimensionality of the desired decomposition, $\mathcal{W}$ is the matrix of basis vectors, $\mathcal{H}$ is the matrix of mixture weights, and $\mathcal{P}$ is the matrix of approximations.

The above equation is similar to equation (2) but with important differences. In equation (2), the matrix undergoing decomposition is $\mathbf{V}$ whereas the goal here is to decompose the original data matrix $\mathbf{V}$. The

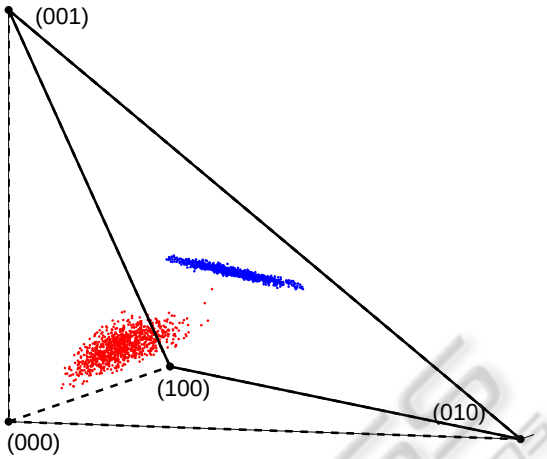

Figure 2: Illustration of normalization on a dataset. Points in red represents a dataset of 1000 3-dimensional points where the directions of maximum variance are orthogonal to the plane corresponding to the standard simplex. Thus, the projection of points in the dataset onto the standard simplex removes important information about the distribution of datapoints.

matrix $\mathcal{W}$ is analogous to $\mathbf{W}$ from equation (2) but unlike the columns of $\mathbf{W}$ that are constrained to sum to 1, the columns of $\mathcal{W}$ have no such constraints. Similarly, $\mathcal{P}$ is analogous to $\mathbf{P}$ but the columns of the former are not constrained to sum to 1 like the columns of $\mathbf{P}$. However, since both equations (2) and (3) are simplex decompositions, matrices $\mathcal{H}$ and $\mathbf{H}$ are alike with entries in each of their columns constrained to sum to 1.

3.3 Algorithm

Consider a scenario where a dataset $\mathbf{V}$ that we desire to decompose using PLSA already lies on the standard simplex. Then, all the constraints that we need as described in the previous subsection are already satisfied. Since all data points lie on the standard simplex, the dataset $\mathbf{V}$ is identical to its normalized version $\bar{\mathbf{V}}$. Hence, the decomposition desired in equation (3) becomes identical to the decomposition in equation (2). We can apply PLSA directly to the given dataset $\mathbf{V}$ and obtain the desired basis vectors.

This observation points to the approach we present below. If we could transform the dataset so that all points lie on the standard simplex and the transformation is invertible, we can achieve the desired decomposition. However, the standard simplex in M-dimensional space represents part of the $(M-1)$-dimensional hyperplane. Thus, instead of being able to have the points exactly lie on the standard simplex, we are constrained to transforming data such that the projections of the data onto $(M-1)$ dimensions of

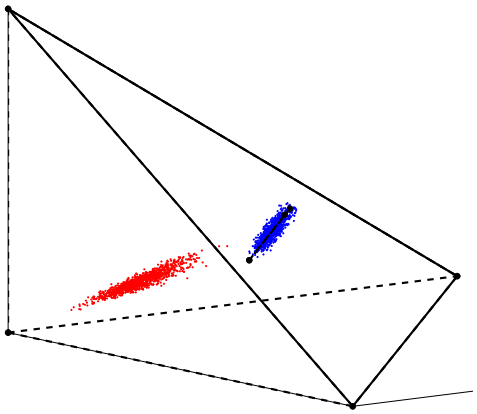

Figure 3: Results of applying PLSA on the dataset shown in Figure 2. Since the projections of data points on the standard simplex (shown in blue) have very narrow variance in one direction, the PLSA simplex obtained is degenerate and almost forms a straight line through the data projections.

our choice will lie on the simplex. Choosing the first $(M-1)$ principal components of the dataset as the $(M-1)$ dimensions on which data will be projected will produce the least error of all possible projections.

The problem now reduces to finding the right transformation that takes the projections of the data on the first $(M-1)$ principal components and aligns them parallel to the standard simplex. The last principal component is transformed such that it left orthogonal to the standard simplex. We leverage the work of (Shashanka, 2009) to define this transformation matrix. More details can be found in the Appendix.

Given the $M \times N$ data matrix $\mathbf{V}$, the entire algorithm can be summarized as follows:

1. Center the data by removing the mean vector to obtain $\hat{\mathbf{V}}$, i.e. $\hat{\mathbf{V}} = \mathbf{V} - mean(\mathbf{V})$.
2. Perform SVD of matrix $\hat{\mathbf{V}}^T$ to obtain $\mathbf{U}$, the matrix of data projections on the singular vectors, i.e. $\hat{\mathbf{V}}^T = \mathbf{USX}^T$.
3. Obtain the $M \times M$ transformation matrix $\mathbf{T}$ (see Appendix for details of this computation).
4. Transform the data to lie parallel to the standard simplex, i.e. $\mathbf{B} = (\mathbf{UT}^T)^T$.
5. Center the transformed data such that the centroid of the simplex coincides with the data mean, i.e. $\bar{\mathbf{B}} = \mathbf{B} - mean(\mathbf{B}) + \mathbf{c}$, where $\mathbf{c}$ is a vector corresponding to the centroid of the standard simplex.
6. Ensure all entries of $\bar{\mathbf{B}}$ are nonnegative by subtracting the minimum entry from the matrix, i.e. $\hat{\mathbf{B}} = \bar{\mathbf{B}} - min(\bar{\mathbf{B}})$.
7. Normalize the matrix $\hat{\mathbf{B}}$ such that entries of the center of the dataset sum to 1, i.e. $\mathbf{B}' = \hat{\mathbf{B}}/b$, where $b = 1 - min(\bar{\mathbf{B}})$.
8. The matrix is now ready for PLSA. Apply PLSA on $\mathbf{B}'$ to obtain $\mathbf{W}$ and $\mathcal{H}$, i.e. $\mathbf{B}' \approx \mathbf{W}\mathcal{H}$.
9. Undo steps 7, 6, 5 and 4 respectively for the basis vector matrix $\mathbf{W}$ to obtain $\bar{\mathbf{W}}$, i.e.
   - $\mathbf{W} = \mathbf{W} \times b$
   - $\mathbf{W} = \mathbf{W} + min(\bar{\mathbf{B}})$
   - $\mathbf{W} = \mathbf{W} + mean(\mathbf{B}) - \mathbf{c}$
   - $\bar{\mathbf{W}} = \mathbf{W}^T\mathbf{T}$
10. Undo the SVD projection and data centering for $\bar{\mathbf{W}}$ to obtain $\mathcal{W}$, i.e.
    - $\bar{\mathbf{W}} = (\bar{\mathbf{W}}\mathbf{SX}^T)^T$
    - $\mathcal{W} = \bar{\mathbf{W}} + mean(\mathbf{V})$

The desired decomposition is given by $\mathbf{V} \approx \mathcal{W}\,\mathcal{H}$.

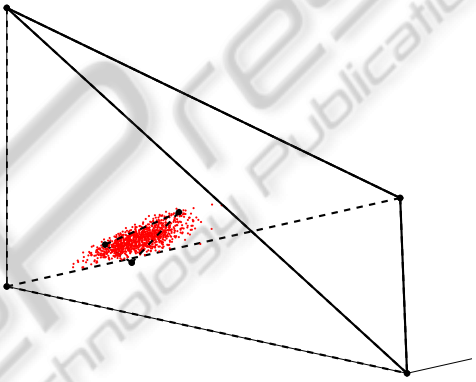

Figure 4: Result of applying our approach on the dataset illlustrated in Figure 2. As desired, the extracted basis vectors form a simplex (dotted black line around the red points) around the original datapoints instead of around the data projections on the standard simplex.

For experiments, we created a synthetic dataset of 1000 3-dimensional points as illustrated in Figure 2. The dataset was created in such a way that the directions of maximal variance present in the data was orthogonal to the plane of the standard simplex. Results of applying PLSA on the dataset is summarized in Figure 3 and results of applying the proposed approach is illustrated in Figure 4.

3.4 Discussion

We first point out that even though we have used PLSA as the specific example, the proposed approach is applicable to any topic modeling technique such as Latent Dirichlet Allocation or Correlated Topic Models where data distributions are expressed as linear combinations of charateristic basis distributions.

In the approach described in the previous subsections, no explicit constraints were placed as to the nonnegativity of the entries of basis vectors. So far in

this paper, we have focused on data that have nonnegative entries but the proposed approach is also applicable for datasets with real-valued entries. The algorithm described earlier can be applied to any arbitrary datasets with real-values entries without any modifications. This is an alternative approach to the one proposed by (Shashanka, 2009). In that work, data is transformed into the next higher dimension so that PLSA can be applied while in this work, we use SVD to align the dataset along the dimensions of the standard simplex. It will be instructive to compare the two approaches in this context and we leave that for future work.

4 CONCLUSIONS

In this paper, we presented a novel approach to perform Simplex Decompositions on datasets. Specifically, the approach learns a set of basis vectors such that each data vector can be expressed as a linear combination of the learned set of bases and where the corresponding mixture weights are nonnegative and sum to 1. PLSA performs a similar decomposition but it characterizes the normalized datapoints instead of the original dataset itself. We demonstrated the spurious effect such a normalization can have with the help of a synthetic dataset. We described our approach and demonstrated that it provides a way to overcome this drawback. This work has several potential applications in tasks such as clustering, feature extraction, and classification. We would like to continue this work by applying the technique on real-world problems and demonstrating its usefulness. We also intend to extend this work to be applicable to other related latent variable methods such as Probabilistic Latent Component Analysis.

REFERENCES

Blei, D. and Lafferty, J. (2006). Correlated Topic Models. In *NIPS*.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet Allocation. *Jrnl of Machine Learning Res.*, 3.

Gaussier, E. and Goutte, C. (2005). Relation between PLSA and NMF and Implications. In *Proc. ACM SIGIR Conf. on Research and Dev. in Information Retrieval*, pages 601–602.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42.

Lee, D. and Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401.

Lee, D. and Seung, H. (2001). Algorithms for Non-negative Matrix Factorization. In *NIPS*.

Shashanka, M. (2009). Simplex Decompositions for Real-Valued Datasets. In *Proc. Intl. Workshop on Machine Learning and Signal Processing*.

Shashanka, M., Raj, B., and Smaragdis, P. (2008). Probabilistic latent variable models as non-negative factorizations. *Computational Intelligence and Neuroscience*.

APPENDIX

In this appendix, we briefly describe how to choose the transformation matrix $\mathbf{T}$ that transforms M-dimensional data $\mathbf{V}$ such that the first $(M-1)$ principal components lie parallel to the standard $(M-1)$-Simplex. We need to indentify a set of $(M-1)$ M-dimensional orthonormal vectors that span the standard $(M-1)$-simplex.

(Shashanka, 2009) developed a procedure to find exactly such a matrix and the method is based on induction. Let $\mathbf{R}_M$ denote a $M \times (M-1)$ matrix of $(M-1)$ orthogonal vectors. Let $\vec{\mathbf{1}}_M$ and $\vec{\mathbf{0}}_M$ denote M-vectors where all the entries are 1's and 0's respectively. Similarly, let $\mathbf{1}_{a\times b}$ and $\mathbf{0}_{a\times b}$ denote $a \times b$ matrices of all 1's and 0's respectively. They showed that the matrix $\mathbf{R}_{(M+1)}$ given by

$$\begin{bmatrix} \mathbf{R}_M & \vec{\mathbf{1}}_M \\ \vec{\mathbf{0}}^T_{(M-1)} & -M \end{bmatrix} \quad \text{if } M \text{ is even, and}$$

$$\begin{bmatrix} \mathbf{R}_{(M+1)/2} & \mathbf{0}_{(M+1)/2\times(M-1)/2} & \vec{\mathbf{1}}_{(M+1)/2} \\ \mathbf{0}_{(M+1)/2\times(M-1)/2} & \mathbf{R}_{(M+1)/2} & -\vec{\mathbf{1}}_{(M+1)/2} \end{bmatrix},$$

if M is odd, is orthogonal. $\mathbf{R}_{(M+1)}$ is then normalized to obtain an orthonormal matrix.

Given the above relation and the fact that $\mathbf{R}_1$ is an empty matrix, one can compute $\mathbf{R}_M$ inductively for any value of M.

We have an additional constraint that the last principal component be orthogonal to the standard simplex and this can be easily achieved by appending a column vector of 1's to $\mathbf{R}_M$.

Thus, the matrix $\mathbf{T}$ defining our desired transformation is given by $[\mathbf{R}_M \quad \vec{\mathbf{1}}_M]$.