

Relation between PLSA and NMF and Implications

Eric Gaussier
Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan, France
Eric.Gaussier@xrce.xerox.com

Cyril Goutte
Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan, France
Cyril.Goutte@xrce.xerox.com

ABSTRACT

Non-negative Matrix Factorization (NMF, [5]) and Probabilistic Latent Semantic Analysis (PLSA, [4]) have been successfully applied to a number of text analysis tasks such as document clustering. Despite their different inspirations, both methods are instances of *multinomial PCA* [1]. We further explore this relationship and first show that PLSA solves the problem of NMF with KL divergence, and then explore the implications of this relationship.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Clustering;
I.5.3 [Clustering]: Algorithms

General Terms

Algorithms

Keywords

Document clustering, probabilistic models, PLSA, NMF

1. INTRODUCTION

Non-negative Matrix Factorization (NMF, [5]) decomposes a (positive) matrix \mathbf{V} into a product of non-negative factors:

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \quad (1)$$

It has been shown that NMF offers a number of advantages in various contexts, in particular when data must be decomposed into a sum of additive components. NMF has been used for example to cluster (textual) documents [5, 8].

Probabilistic Latent Semantic Analysis (PLSA, [4]) is a model-based document clustering technique. In a collection of J documents indexed with I words, PLSA models the $I \times J$ term-document co-occurrence matrix \mathbf{M} (where M_{ij} is the number of occurrences of word w_i in document d_j) as arising from a mixture model with K components:

$$P(w_i, d_j) = \sum_{c=1}^K P(c)P(d_j|c)P(w_i|c) \quad (2)$$

Parameters are estimated by maximizing the likelihood of the observed data \mathbf{M} .

Earlier studies [1] showed that NMF and PLSA are both instances of a more general model, and are therefore linked. In this short note, we show a formal equivalence between these methods, down to almost identical update rules, and discuss its implications.

2. PLSA IS NMF WITH KL DIVERGENCE

Introducing the $I \times K$ matrix \mathbf{W}' s.t. $W'_{ic} = P(w_i|c)P(c)$, and the $K \times J$ matrix \mathbf{H}' s.t. $H'_{cj} = P(d_j|c)$, we can rewrite eq. 2 as $[P(w_i, d_j)] = \mathbf{W}'\mathbf{H}'$. As probability matrices are obviously non-negative, PLSA corresponds to factorizing the joint probability matrix in non-negative factors.

Conversely, given the NMF formulation in eq. 1, assume without loss of generality that $\sum_{ij} V_{ij} = 1$ (otherwise \mathbf{V} , \mathbf{W} and \mathbf{H} may be appropriately scaled). Let us introduce \mathbf{A} and \mathbf{B} , two $K \times K$ diagonal scaling matrices such that $A_{kk} = \sum_i W_{ik}$ and $B_{kk} = \sum_j H_{kj}$. We have:

$$\mathbf{W}\mathbf{H} = (\mathbf{W}\mathbf{A}^{-1}\mathbf{A})(\mathbf{B}\mathbf{B}^{-1}\mathbf{H}) = (\mathbf{W}\mathbf{A}^{-1}) \times (\mathbf{A}\mathbf{B}) \times (\mathbf{B}^{-1}\mathbf{H})$$

where $(\mathbf{W}\mathbf{A}^{-1})$ (resp. $(\mathbf{B}^{-1}\mathbf{H})$) has all the formal properties of the conditional probability matrix $[P(w_i|c)]$ (resp. $[P(d_j|c)]^\top$), and $(\mathbf{A}\mathbf{B})$ is akin to a diagonal matrix of $P(c)$. In passing, we note that the NMF is invariant if we multiply \mathbf{W} by a diagonal matrix and \mathbf{H} by its inverse. The following property expresses a stricter relationship between PLSA and NMF.

Property Any (local) maximum likelihood solution of PLSA is a solution of NMF with KL divergence.

Proof. Without loss of generality, we assume that the positive matrix to be decomposed is a co-occurrence matrix¹ \mathbf{M} . Let \mathbf{V} be the scaled matrix $V_{ij} = M_{ij}/N$, $N = \sum_{ij} M_{ij}$. Solutions of NMF with KL divergence are fixed points of the following update rules [5, 6]:

$$H_{cj} \leftarrow H_{cj} \frac{\sum_i \frac{W_{ic} V_{ij}}{(WH)_{ij}}}{\sum_i W_{ic}}, \quad W_{ic} \leftarrow W_{ic} \frac{\sum_j \frac{H_{cj} V_{ij}}{(WH)_{ij}}}{\sum_j H_{cj}} \quad (3)$$

The EM algorithm used to maximize the likelihood in PLSA iterates two steps, and solutions are fixed points of the following equations [4]:

$$P(c)^{(t+1)} = \sum_{i,j} \frac{M_{ij}}{N} P(c|w_i, d_j)^{(t)}$$

¹Any positive matrix can be approximated, wrt the L_1 or L_2 norm, arbitrarily well as the product of a co-occurrence matrix and a scalar, the latter being just a scale factor in the NMF or PLSA decomposition.

$$P(w_i|c)^{(t+1)} = \frac{\sum_j M_{ij} P(c|w_i, d_j)^{(t)}}{\sum_{i,j} M_{ij} P(c|w_i, d_j)^{(t)}}$$

$$P(d_j|c)^{(t+1)} = \frac{\sum_i M_{ij} P(c|w_i, d_j)^{(t)}}{\sum_{i,j} M_{ij} P(c|w_i, d_j)^{(t)}}$$

with $P(c|w_i, d_j)^{(t)} = P(w_i, c)^{(t)} P(d_j|c)^{(t)} / P(w_i, d_j)^{(t)}$.

Joining the equations for $P(c)$ and $P(d_j|c)$, we obtain $P(d_j|c)^{(t+1)} = (\sum_i M_{ij} P(c|w_i, d_j)^{(t)}) / (N \cdot P(c)^{(t+1)})$. Expanding $P(c|w_i, d_j)^{(t)}$ and using the notation introduced above:

$$H'_{cj}{}^{(t+1)} = H'_{cj}{}^{(t)} \frac{\sum_i \frac{W'_{ic}{}^{(t)} V_{ij}}{(W'_{ic}{}^{(t)} H'_{ij}{}^{(t)})}}{\sum_i W'_{ic}{}^{(t+1)}}$$

Similarly, one obtains:

$$W'_{ic}{}^{(t+1)} = W'_{ic}{}^{(t)} \sum_j \frac{H'_{cj}{}^{(t)} V_{ij}}{(W'_{ic}{}^{(t)} H'_{ij}{}^{(t)})} = W'_{ic}{}^{(t)} \frac{\sum_j \frac{H'_{cj}{}^{(t)} V_{ij}}{(W'_{ic}{}^{(t)} H'_{ij}{}^{(t)})}}{\sum_j H'_{cj}{}^{(t)}}$$

At any fixed point of the above equations, $W'_{ic}{}^{(t+1)} = W'_{ic}{}^{(t)}$. Thus, any fixed point of EM is also a fixed point of the update rules 3. \square

Using a similar fixed point argument, as well as the formal equivalence above, it is possible to show that the converse is true, that is: Any solution of NMF with KL divergence yields a (local) maximum likelihood solution of PLSA.

3. IMPLICATIONS

The relation we demonstrated in the previous section has a number of consequences. The main implication is that whenever a problem may be formulated with NMF (for example word alignment, [3]), it may be efficiently solved using PLSA. There are in fact many advantages to do so. First, PLSA is a probabilistic model which offers the convenience of the highly consistent probabilistic framework. Whereas the NMF factors are a set of values (with scale invariance issues, cf. below), the PLSA parameters may be interpreted as probabilities. We may for example evaluate the importance of the factors ($P(c)$) and interpret each factor as a probabilistic profile ($P(w|c)$ or $P(d|c)$). The probabilistic framework also comes in handy in some situations such as orthogonalizing the non-negative factors (for example [3]).

Regarding parameter estimation, the above development shows that the NMF update rules are essentially an EM algorithm. This means that they must be prone to the usual problems associated with EM such as sensitivity to local minima², even though, to our knowledge, this is not widely acknowledged in the NMF literature. On the other hand, with probabilistic models, it is possible to benefit from the considerable body of work dedicated to addressing various shortcomings of EM. Such advances include the use of “tempered” (or deterministic annealing) EM [4] in order to stabilize the parameter estimation, reduce the sensitivity to initial conditions, and even, to some extent, optimize the number of components in the mixture. A related aspect is the selection of a proper model structure. Many techniques have been proposed in the literature for assessing the number of components in a mixture model [7]. In comparison,

²Indeed, the property established before implies that NMF has at least as many local optima as PLSA.

the NMF literature puts little emphasis on the choice of the correct number of factors.

We noted in passing that a NMF is invariant if we multiply \mathbf{W} and \mathbf{H} by a square matrix and its inverse, respectively. This shows that there are infinitely many equivalent factorizations (in addition to the permutations of factors). In the vocabulary of mixture models, the NMF factors are not identifiable, whereas the PLSA model is. This has implications on theoretical properties of the Maximum Likelihood estimator of PLSA, such as strong consistency [7]. In addition, let us note that PLSA has hierarchical extensions [2]. These are especially convenient for document clustering (and categorization), because it is quite common to organize documents in hierarchies of classes. NMF offers no such extension and is therefore limited to “flat” factors.

However, as NMF is expressed in terms of matrix approximation, different results may be obtained by optimizing different measures of the approximation. In fact, [5] derives update rules corresponding to a squared (Euclidean) approximation loss. We are not aware of any direct equivalence to this in terms of Maximum Likelihood estimation with PLSA, and we are currently working on adapting the PLSA estimation model to different losses (squared loss, absolute value, etc.). Finally, the relation we have exhibited also supports the results reported in document clustering using NMF (for example [8]). In particular, as PLSA has been designed to overcome some limitations of LSI, it is not surprising that NMF, being essentially equivalent, also outperforms LSI.

4. CONCLUSION

This short note exhibits a remarkable relation between two common document clustering techniques, NMF and PLSA, and shows that PLSA solves NMF with KL divergence. This relation allows us to solve the NMF problem with a probabilistic mixture model for which various extensions and technical advances have been proposed.

Acknowledgement

We thank François Yvon for discussions related to this topic.

5. REFERENCES

- [1] W. Buntine. Variational extensions to EM and multinomial PCA. In *ECML'02*, 2002.
- [2] E. Gaussier, C. Goutte, K. Popat, and F. Chen. A hierarchical model for clustering and categorising documents. In *Advances in Information Retrieval*, Lecture Notes in Computer Science. Springer, 2002.
- [3] C. Goutte, K. Yamada, and E. Gaussier. Aligning words using matrix factorisation. In *ACL'04*, 2004.
- [4] T. Hofmann. Probabilistic latent semantic analysis. In *UAI'99*, pages 289–296. Morgan Kaufmann, 1999.
- [5] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.
- [6] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS'13*. MIT Press, 2001.
- [7] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.
- [8] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *SIGIR'03*, 2003.