The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009
Copyright © 2009 ORSC & APORC, pp. 41–47

# NMF-based Models for Tumor Clustering: A Systematic Comparison

Zhong-Yuan Zhang[1,*]

[1]School of Statistics, Central University of Finance and Economics, Beijing 100080

**Abstract** Nonnegative Matrix Factorization (NMF) is one of the famous unsupervised learning models. In this paper, we give a short survey on NMF-related models, including K-means, Probabilistic Latent Semantic Indexing etc. and present a new Posterior Probabilistic Clustering model, and compare their numerical experimental results on five real microarray data. The results show that i) NMF using with K-L divergence objective function has better clustering performance; ii) Our purposed PPC model is among the best.

**Keywords** Algorithm; Nonnegative Matrix Factorization; Microarray; Comparison; Bioinformatics

# 1 Introduction

Nonnegative Matrix Factorization (NMF) is one of unsupervised learning models used for data mining. Generally speaking, NMF factorizes some nonnegative matrix $X$, which often comes from biology, text or image, into two nonnegative matrices that satisfy $X \approx FG^T$. NMF has been successfully applied for data clustering, dimensional reduction, image processing, etc. [15, 7, 1, 2, 13].

There are many papers that are devoted to analyze NMF from different perspectives. In them, several variations of NMF are given and studied, and the equivalence between NMF and other classical unsupervised learning models, such as K-means and Probabilistic Latent Semantic Indexing, are proved [3, 4]. In this submission, we give a short survey on five NMF-related models. Furthermore, we also present a new variation of NMF, called Posterior Probabilistic Clustering (PPC). Experiment results show that PPC can give better results. In summary, The contributions of this submission are three folds: (1) A new posterior probabilistic clustering model is presented; (2) A short survey on different NMF-related models is given; (3) A systematic comparison on the different models used for microarray data is given. Results show that our purposed PPC model is valuable and competitive.

The rest of the paper is organized as follows: Section 2 is a survey on different NMF-related models; Section 3 gives a systematic comparison of the different models on five microarray datasets. Section 4 concludes and discusses the future work.

[*] E-mail: zhyuanzh@gmail.com

## 2 A Short Survey on NMF-related Models for Tumor Clustering

In this section, we give a brief survey on NMF-related models, including K-means, Probabilistic Latent Semantic Indexing, nsNMF. We also give a new model called Posterior Probabilistic Clustering.

**Nonnegative Matrix Factorization, NMF ([9, 10]):** In general, NMF can be written as:

$$\begin{array}{ll} \min & J(X, FG^T) \\ s.t. & F, S, G \geqslant 0. \end{array}$$

$J(X, FG^T)$ is some distance function or dissimilarity function between two matrices $X$ and $FG^T$. $F$ and $G$ are updated alternately until convergence.

If the least square error $\|X - FG^T\|_F^2$ is selected as objective function $J$ to optimize, the corresponding update rules of $F$ and $G$ are:

$$\begin{aligned} F_{ia} &:= F_{ia}\frac{(XG)_{ia}}{(FG^TG)_{ia}} \\ G_{ia} &:= G_{ia}\frac{(X^TF)_{ia}}{(GF^TF)_{ia}} \end{aligned}$$

Otherwise, if the K-L divergence $\sum\limits_{i,j}(X_{ij}\log\frac{X_{ij}}{(FG^T)_{ij}} - X_{ij} + (FG^T)_{ij})$ is selected to optimize, the rules of $F$ and $G$ are:

$$F_{ia} := \frac{F_{ia}}{\sum\limits_j G_{ja}}\sum_j \frac{X_{ij}}{(FG^T)_{ij}}G_{ja} \tag{1}$$

$$G_{ja} := \frac{G_{ja}}{\sum\limits_i F_{ia}}\sum_i \frac{X_{ij}}{(FG^T)_{ij}}F_{ia} \tag{2}$$

**K-means:** [3] shows that NMF that factorizes symmetric matrix $X$, which is the similarity matrix of the original samples, with orthogonal constraints on the factor matrices $F$ (or $G$) is equivalent to K-means.

**Probabilistic Latent Semantic Indexing, PLSI ([6]):** PLSI is one of the topic models that are successfully applied to information retrieval. [4] shows that PLSI and NMF optimize the same objective function (K-L divergence) while use different update rules. If $X$ is normalized to satisfy $\sum\limits_{ij} X_{ij} = 1$, the model can be written as:

$$\min \quad \sum_{i,j}(X_{ij}\log\frac{X_{ij}}{(FSG^T)_{ij}} - X_{ij} + (FSG^T)_{ij}) \tag{3}$$

$$s.t. \quad F, S, G \geqslant 0 \tag{4}$$

$$\sum_i F_{ik} = 1, \sum_j G_{jk} = 1, \sum_k S_{kk} = 1. \tag{5}$$

$S$ is diagonal. Our results show that the update rules of $F$ and $G$ in PLSI are indeed got from NMF simply by normalizing $F$ and $G$ in Eq.(1),(2) at each iteration. Details will come in our future work.

**Nonsmooth NMF, nsNMF ([11]):** nsNMF optimizes $X = FSG^T$ instead of $X = FG^T$, where $S = (1-\theta)I + \dfrac{\theta}{k} II^T$, $I$ is the identity matrix and the parameter $\theta$ is used to control the sparseness of both $F$ and $G$. More details can be found in [11].

**Posterior Probabilistic Clustering:** Different from PLSI, which treats the factor matrices $F$, $S$ and $G$ as class-conditional probabilistic matrices, i.e., $F$, $S$ and $G$ satisfy the condition (5), PPC ([5]) regards $F$, $S$ and $G$ as posterior probabilistic matrices, i.e., $\sum\limits_k F_{ik} = 1, \sum\limits_k G_{jk} = 1, \sum\limits_k S_{kk} = 1$. To simplify the model, we only add constraint on $G$. Different from [5], we select K-L divergence as objective function and we briefly give the algorithms of PPC. Details of motivations of PPC and convergence analysis will come soon in our future work.

The model can be written as:

$$\begin{array}{lll} (PPC) & \min\limits_{F,G\geqslant 0} & \sum_{i,j}(X_{ij} \log \frac{X_{ij}}{(FG^T)_{ij}} - X_{ij} + (FG^T)_{ij}) \\ & s.t. & \sum\limits_{k=1}^{K} G_{jk} = 1, \quad j = 1,2,\cdots,n. \end{array}$$

Mimic the derivative process of PLSI, we can get the update rules of $F$ and $G$.

$$\begin{aligned} G_{jk} &= \frac{G_{jk}}{\sum\limits_j F_{jk} \sum\limits_k G_{jk}} \sum_i \frac{X_{ij} F_{ik}}{(FG^T)_{ij}} \\ &= \frac{G_{jk}(\dfrac{X^T F}{GF^T})_{jk}}{\sum\limits_i F_{ik} \sum\limits_k [\dfrac{G_{jk}(\frac{X^T F}{GF^T})_{jk}}{\sum\limits_i F_{ik}}]}, \end{aligned}$$

and the update rule of $F$ is the same as the standard NMF:

$$F_{ik} \quad := \quad \frac{F_{ik}}{\sum\limits_j G_{jk}} (\frac{X}{FG^T} G)_{ik}$$

Details will also come in our future work.

Note that a detailed analysis of PPC using with Lease Squares Error will also come soon in our future work. For short, we employ penalty function algorithm to solve the model. The numerical results are listed in Table 1 and Table 2

# 3 Experiment on Microarray Data

In this section, we give a systematic comparison of different models on five microarray datasets. The results show that our proposed PPC model is competitive.

### 3.1 Data Description

We use five datasets to assess the performance of the six models. Note that "Subtypes of Acute Lymphoblastic Leukemia" includes two datasets "BCR-ABL/E2A-PBX1/MLL" and "Hyperdip50/MLL/T-ALL"

**ALL-AML**This dataset includes two types of human tumor-acute myelogenous leukemia (AML, 11 samples) and acute lymphoblastic leukemia (ALL, 27 samples). Also ALL can be divided into two subtypes-ALL-T(8 samples) and ALL-B(19 samples).[1]

**Central Nervous System(CNS)** This dataset comes from [12] which consists of 34 samples: 10 classic medulloblastomas, 10 malignant, gliomas, 10 rhabdoids and 4 normals.

**Lung cancer (LC)** This dataset, composed of 32 samples, is from [8] which is about malignant pleural mesothelioma (MPM, 16 samples) and adenocarcinoma (ADCA, 16 samples) of the lung.

**Subtypes of Acute Lymphoblastic Leukemia:** This dataset includes six prognostically important leukemia subtypes: T-ALL, E2A-PBX1, BCR-ABL, TEL-AML1, MLL, hyperdiploid>50 chromosomes. We select E2A-PBX1 (18 samples), MLL (14 samples), T-ALL (28 samples) as one test dataset, and E2A-PBX1 (18 samples), Hyperdiploid>50 (42 samples), T-ALL (28 samples), TEL-AML1 (52 samples) as another. The original data contains about 12000 genes. In our experiment, the genes are ranked according to their coefficient of variation (i.e., standard deviation divided by the mean) and the top 5000 are selected.

### 3.2 Experiment Results

In order to compare the clustering performance, we use Normalized Mutual Information (NMI) and Accuracy(ACC) as our performance measures.

NMI is computed as:

$$NMI = \frac{\sum_{i,j} P(i,j) log_2 \frac{P(i,j)}{P(i)P(j)}}{\sqrt{(\sum_i -P(i)log_2 P(i))(\sum_j -P(j)log_2 P(j))}}, \tag{6}$$

where $P(i)$ is the probability that an arbitrary data point belongs to computed class $i$, and $P(j)$ is the probability that an arbitrary data point belongs to implanted class $j$. $P(i,j)$ is the joint probability that an arbitrary data point belongs to cluster $i$ and also class $j$. Details can be found at [14]. Generally, the larger the NMI value, the better the clustering quality is. Its value is between 0 and 1.

Accuracy is computed as:

$$ACC = \max(\sum_{C_k, L_m} T(C_k, L_m))/N, \tag{7}$$

where $C_k$ is the $k$-th computed class, and $L_m$ is the $m$-th ground-truth class. $T(C_k, L_m)$ is the number of samples that belong to class $m$ are assigned to cluster $k$. Generally, the larger the accuracy value, the better the clustering performance.

All the algorithms are performed by MATLAB 7.6.0.324. The parameter $\theta$ in nsNMF is 0.5, as recommended by the original paper [11]. All the results are averages of ten runs of the algorithms.

Table 1: Accuracy Comparisons on various datasets. Each entry is the percentage of clustering accuracy of the column method on the corresponding row dataset. "NMF1" denotes "NMF using with least squares error" and "NMF2" denotes "NMF using with K-L divergence". "PPC1" denotes "PPC using with least squares error" and "PPC2" denotes "PPC using with K-L divergence". The bold ones are the best results that are got from methods based on K-L divergence, while the underlined ones are the best results got from least squares error.

| | K-means | NMF1 | NMF2 | PLSI | nsNMF | PPC1 | PPC2 |
|---|---|---|---|---|---|---|---|
| Multi-class I | 64.63 | 72.20 | **75.37** | 68.05 | 61.71 | 74.15 | **75.37** |
| Multi-class II | 57.86 | 66.43 | 72.50 | **77.86** | 72.98 | 72.50 | 72.38 |
| Lung Cancer | 76.56 | 93.75 | **100.00** | **100.00** | 100.00$\surd$ | 81.25 | **100.00** |
| AML/ALL2 | 74.47 | 96.58 | 95.00 | 94.74 | 94.74 | 80.00 | **97.37** |
| AML/ALL3 | 89.21 | 95.26 | **95.53** | 92.11 | 94.74 | 77.63 | 94.74 |
| CNS | 77.06 | 96.47$\surd$ | 94.41 | 94.12 | 93.24 | 93.82 | 94.12 |

Table 2: NMI Comparisons on various datasets. Each entry is the percentage of clustering NMI of the column method on the corresponding row dataset. "NMF1", "NMF2", "PPC1" and "PPC2" have the same meanings in Tab 1. The bold ones and the underlined ones also have the same meanings in Tab 1.

| | K-means | NMF1 | NMF2 | PLSI | nsNMF | PPC1 | PPC2 |
|---|---|---|---|---|---|---|---|
| Multi-class I | 42.58 | 44.49 | **46.20** | 37.13 | 28.23 | 47.32 | 45.99 |
| Multi-class II | 12.78 | 36.87 | 53.59 | **55.51** | 44.41 | 38.62 | 49.11 |
| Lung Cancer | 34.12 | 71.69 | **100.00** | **100.00** | 100.00$\surd$ | 42.75 | **100.00** |
| AML/ALL2 | 33.87 | 77.79 | 71.91 | 70.80 | 70.80 | 21.33 | **81.13** |
| AML/ALL3 | 75.42 | 83.25 | **83.61** | 74.98 | 80.78 | 39.84 | 81.59 |
| CNS | 66.57 | 91.08$\surd$ | 86.55 | 85.87 | 89.38 | 84.81 | 85.20 |

Table 1 and Table 2 demonstrate the accuracy and NMI comparison results of different methods, from which we can observe that: (1) K-means is always the worst; (2) for microarray datasets, the methods that use K-L divergence as objective function are better than that use least squares error almost all the times (CNS is the only exception); (3) NMF2 and PPC2 are the winners. For three out of six dataset, they give the best results.

## 4 Conclusion and Future Works

In this paper, we give a short survey including six models for regulations on NMF. We present a posterior probability clustering model. We also give a systematic comparison of the six models on five real-world datasets coming from biology society. The results show that our proposed model PPC is a valuable and competitive one among the six models. Hence a further study of PPC is of interest.

As to future work, we can generalize PPC to simultaneous feature and sample clustering. We use $F$ as the posterior probability for feature clustering, and the posterior probability normalization is $\sum_{k=1}^{K} F_{ik} = 1$. The simultaneous PPC (SPPC) becomes

$$\begin{aligned} \min_{F,S,G \geqslant 0} \quad & J(X, FSG^T), \\ s.t. \quad & \sum_{k=1}^{K} F_{ik} = 1, \sum_{k=1}^{K} G_{jk} = 1, \end{aligned}$$

where $J(X, FSG^T)$ can be conventional least squares error or K-L divergence, the corresponding algorithms can be derived similarly to PPC.

For example, if K-L divergence is selected, the update rules of $F, S$ and $G$ are:

$$\begin{aligned} G_{jk} \quad &:= \quad G_{jk} \frac{(\frac{X^T}{GSF^T} FS)_{jk}}{\sum_k G_{jk} (\frac{X^T}{GSF^T} FS)_{jk}} \\ &= \quad G_{jk} \frac{(\frac{X^T}{GSF^T} FS)_{jk}}{(GSF^T \frac{X}{FSG^T})_{jj}} \\ F_{jk} \quad &:= \quad F_{jk} \frac{(\frac{X}{FSG^T} GS)_{jk}}{(FSG^T \frac{X^T}{GSF^T})_{jj}} \\ S_{kk} \quad &:= \quad S_{kk} (F^T \frac{X}{FSG^T} G)_{kk}. \end{aligned}$$

## Acknowledges

This work is supported by the National Natural Science Foundation of China under grant No.60873205 and Foundation of Academic Discipline Program, Central University of Finance and Economics.

# References

[1] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, 101(12):4164–4169, March 2004.

[2] Karthik Devarajan. Nonnegative matrix factorization: An analytical and interpretive tool in computational biology. *PLoS Comput Biol*, 4(7):e1000029+, July 2008.

[3] C. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SIAM Data Mining Conf*, pages 606–610, 2005.

[4] C. Ding, T. Li, and W. Peng. Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method. *Proceedings of the National Conference on Artificial Intelligence*, 21(1):342, 2006.

[5] Chris Ding, Tao Li, Dijun Luo, and Wei Peng. Posterior probabilistic clustering using nmf. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 831–832, New York, NY, USA, 2008. ACM.

[6] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM Press, 1999.

[7] P. O. Hoyer. Non-negative sparse coding. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pages 557–565, 2002.

[8] Gordon G. J, Jensen R. V, Hsiao L. L, Gullans S. R, Blumenstock J. E, Ramaswamy S, Richards W. G, Sugarbaker D. J, and Bueno R. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res.*, 62(17):4963–7, September 2002.

[9] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

[10] Daniel D. Lee and Sebastian H. Seung. Algorithms for non-negative matrix factorization. In *Annual Conference on Neural Information Processing Systems*, pages 556–562, 2000.

[11] A. Pascual-Montano, J. M. Carazo, K. Kochi, D. Lehmann, and R. D. Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsnmf). *IEEE transactions on Pattern Analysis and Machine Intelligence*, 28(3):403–415, March 2006.

[12] Scott L. Pomeroy, Pablo Tamayo, Michelle Gaasenbeek, Lisa M. Sturla, Michael Angelo, Margaret E. Mclaughlin, John Y. H. Kim, Liliana C. Goumnerova, Peter M. Black, Ching Lau, Jeffrey C. Allen, David Zagzag, James M. Olson, Tom Curran, Cynthia Wetmore, Jaclyn A. Biegel, Tomaso Poggio, Shayan Mukherjee, Ryan Rifkin, Andrea Califano, Gustavo Stolovitzky, David N. Louis, Jill P. Mesirov, Eric S. Lander, and Todd R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, January 2002.

[13] Farial Shahnaz, Michael W. Berry, Paul V. Pauca, and Robert J. Plemmons. Document clustering using nonnegative matrix factorization. In *Journal on Information Processing and Management*, volume 42, 2004.

[14] Alexander Strehl and Joydeep Ghosh. Cluster ensembles: a knowledge reuse framework for combining partitionings. In *Eighteenth national conference on Artificial intelligence*, pages 93–98, Menlo Park, CA, USA, 2002. American Association for Artificial Intelligence.

[15] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, New York, NY, USA, 2003. ACM Press.