School of Computer Science, UWindsor

Vol. 00 no. 00 xxxx
Pages 1–2

The Non-Negative Matrix Factorization Toolbox in MATLAB for High Dimensional Biological Data

Yifeng Li [1,*] and Alioune Ngom [1]

[1]School of Computer Science, University of Windsor, Windsor, Ontario, Canada, N9B 3P4.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Technical Report No. 11-060

ABSTRACT

Motivation: Non-negative matrix factorization (NMF) has attracted much attention in the machine learning and bioinformatics communities due to its impressing applications. Though there exist some packages implemented in R and other programming languages, they either only provide some optimization algorithms, or focus on a specific application field. There is no complete package for the bioinformatics community to perform various analysis on high dimensional data.

Results: We provide a powerful but convenient MATLAB toolbox including the algorithms of various NMFs and a variety of NMF based functions for analyzing biological data. Through this toolbox, analysis approaches such as clustering, biclustering, feature extraction, feature selection, classification, overcoming missing values, and visualization can be easily done.

Availability: The toolbox is non-commercial and is available at http://cs.uwindsor.ca/~li11112c/nmf.

Contact: li11112c@uwindsor.ca

1 INTRODUCTION

A complex biological phenomenon may be only determined by a few potential factors. Based on this assumption, we have the *non-negative linear model*. Given a data $\boldsymbol{X}$, and suppose the k potential factors are columns of $\boldsymbol{A}$. This model allows that a data point observed by experiment is only a non-negative linear combination of these potential factors. The linear model can be formulated as $\boldsymbol{X} = \boldsymbol{AY}$, where $\boldsymbol{Y} \geq 0$. The task of this model is to find the potential factors $\boldsymbol{A}$ and the non-negative coefficients $\boldsymbol{Y}$.

Some biological data, for example gene expression data, are naturally non-negative. If $\boldsymbol{X}$ is non-negative, the potential factors $\boldsymbol{A}$ should also be naturally non-negative. In this case, we need to add the non-negative constraint on $\boldsymbol{A}$ in the above model. The optimization task now is $\min \|\boldsymbol{X} - \boldsymbol{AY}\|_{\mathrm{F}}^2$, s.t. , $\boldsymbol{A}, \boldsymbol{Y} \geq 0$. This becomes a non-negative matrix factorization (NMF) problem proposed by Lee and Seung, 1999. Approximation solutions can be found by either multiple update rules or alternative non-negative least squares algorithm (Kim and Park, 2007). We develop a NMF MATLAB toolbox to provides optimization algorithms and support all of the non-negative linear models based biological data analysis introduced below.

*to whom correspondence should be addressed

2 METHODS

Sparse-NMF and has been applied to cluster microarray data (Kim and Park, 2007). Sparsity constraint is imposed on $\boldsymbol{Y}$ to reduce non-uniqueness and enhance interpretation. Alternatively, Orthogonal-NMF imposes orthogonality to enhance sparsity (Ding *et al*., 2006). Ding *et al*., 2010 proposed Convex-NMF, where the columns of $\boldsymbol{A}$ are constrained to be the convex combinations of data points in $\boldsymbol{X}$. It has been verified that columns of $\boldsymbol{A}$ obtained by convex-NMF are close to the real cluster centroids. Convex-NMF can be kernelized to Kernel-NMF (Ding *et al*., 2010). Kernel-NMF maps the data points to a higher-dimensional space where data are clustered. We implement the algorithms of the standard NMF and its variants mentioned above.

2.1 Clustering

NMF has been applied for clustering. Given data $\boldsymbol{X}$ with data points in the columns, the idea is that, after applying NMF on $\boldsymbol{X}$, a data point ($\boldsymbol{x}_i$) is a non-negative linear combination of the columns of $\boldsymbol{A}$. The greatest coefficient in the ith column of $\boldsymbol{Y}$ indicates the cluster this sample belongs to. The reason is that if the data points are mainly composed by the same potential factor, they should be in the same group. A potential factor is usually viewed as a cluster centroid. It has been applied for clustering microarray data for cancer class discovery (Brunet *et al*., 2004) and mining biological processes (Kim and Park, 2007). For gene expression data, NMF based biclustering can also be applied to simultaneously group the genes and samples. See (Madeira and Oliveira, 2004) for a survey. Its goal is to find strongly correlated genes over a subset of samples. A subset of such genes and a subset of such samples form a bicluster. We implement these approaches based on the NMFs mentioned above. The biclusters can be visualized using this toolbox. An example is shown in Fig. 1.

2.2 Feature Extraction

Microarray data and mass spectrometry data have tens of thousands of features but only tens or hundreds of samples. This problem either makes many approaches break down, or leads to intolerable computing time. Another issue is that the biological data normally contains noise, which crucially affects the analysis. In cancer research, some researchers believe that only several of biological factors play a crucial role. When we get the data from controls and unhealthy patients, the huge amount of data therefore contains lots of irrelevant and redundant information. NMF can be employed to

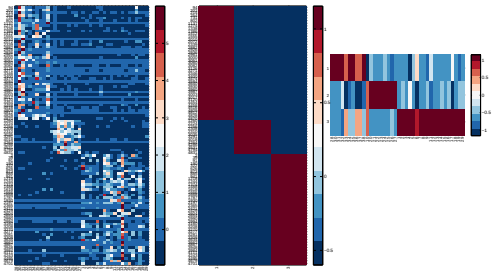

Fig. 1. A biclustering result. The left is a gene expression dataset, the middle is the coefficient matrix, and the right is the coefficient matrix.

extract new features from such data. Training data $\boldsymbol{X}_{m\times n}$, with m features and n samples, can be decomposed into $\boldsymbol{A}_{m\times k}$ and $\boldsymbol{Y}_{k\times n}$, that is $\boldsymbol{X} \approx \boldsymbol{A}\boldsymbol{Y}_{\text{tr}}$, s.t. $\boldsymbol{A}, \boldsymbol{Y}_{\text{tr}} \geq 0$. The k columns of $\boldsymbol{A}$ span the *feature space* and each column of $\boldsymbol{Y}$ is the representation of the corresponding original training sample in the feature space. In order to project the unknown samples $\boldsymbol{S}$ into this feature space, we have to solve the following non-negative least squares problem: $\boldsymbol{S} \approx \boldsymbol{A}\boldsymbol{Y}_{\text{uk}}$, s.t.$\boldsymbol{Y}_{\text{uk}} \geq 0$. Algorithms have been implemented to solve this. After obtaining $\boldsymbol{Y}_{\text{tr}}$ and $\boldsymbol{Y}_{\text{uk}}$, the learning and prediction can be quickly done in the k-dimensional feature space instead of the m-dimensional original space. A classifier can learn over $\boldsymbol{Y}_{\text{tr}}$, and then predict the class labels of unknown sample representations $\boldsymbol{Y}_{\text{uk}}$. Experiments over microarray data have proved that the NMF based feature extraction can achieve higher prediction accuracy than the *principle component analysis* (PCA) (Li and Ngom, 2010).

2.3 Feature Selection

The columns in $\boldsymbol{A}$ for the gene expression data is called *metagenes* (Brunet *et al*., 2004). A metagene can be interpreted as a biological process, because its values imply the activity and silence of all the genes. Gene selection aims to find marker genes to aid the disease prediction and understand the gene regulatory network. Rather than selecting genes on the original data, the novel idea is to conduct gene selection on the metagenes. The reason is that the discovered biological process via NMF are biologically meaningful for the class discrimination, and the genes, expressing differently across these processes, should contribute to the classification. In Fig. 1, for example, three biological processes are discovered and only selected genes are shown. We have implemented the information entropy based gene selection approach (Kim and Park, 2007). It has been reported that it can select meaningful genes, which has been verified by gene ontology analysis.

2.4 Classification

Two novel classifiers based on NMF have be proposed for high dimensional data. The first one is called *non-negative least squares* (NNLS) classifier (Li and Ngom, 2011). The idea is that each unknown sample is assumed to be regressed by a positive linear combination of few training samples, and this unknown sample can be predicted to have the same class label as the training sample with the largest coefficient. Bootstrapping has been used to improve its performance. The second one is a local and transductive learner, termed the NMF classifier (Li, 2011). Firstly, NMF or Semi-NMF is applied to group the union of labeled training samples and unlabeled unknown samples into clusters. Then, for each cluster, the NNLS classifier is employed to predict the class labels of the unknown samples within this cluster. Due to the instability of the NMF, different clustering results may be obtained in different runs with the same input. Therefore, the *repetitive NMF* (RNMF) classifier is devised to overcome this. The idea is to rerun the NMF classifier multiple times and then resort to the majority voting rule. We have implemented these two classification approaches. Experiments have shown that the performance of the NNLS classifiers is better than that of *1-nearest neighbor*, and is comparable with the well-known *support vector machine*. The NMF classifiers outperform them.

2.5 Missing Values

Biological data sometimes suffer from missing values. To deal with this problem, three strategies, removal, imputation, and disregard, are usually used. The first one will remove much other useful information, particularly when there is a large percent of missing values. The second one has a high risk of introducing false data. The third one is to avoid using missing values during analysis, which is the best way. We have implemented the Weighted-NMF (Ho, 2008) to ignore the missing values during computation.

3 CONCLUSION

We propose the powerful NMF MATLAB Toolbox which can be used to analyze high dimensional biological data via clustering, biclustering, feature extraction, feature selection, classification. Missing values can also be avoided during NMF optimization.

ACKNOWLEDGEMENT

Funding: This research has been supported by IEEE CIS Walter Karplus Summer Research Grant 2010, and Canadian NSERC Grants #RGPIN228117-2011.

REFERENCES

Brunet,J.P., Tamayo,P., Golub,T.R., and Mesirov,J.P. (2004) "Metagenes and molecular pattern discovery using matrix factorization," *PNAS*, **101(12)**, 4164-4169.

Ding,C., Li,T., Peng,W., and Park,H. (2006) Orthogonal nonnegative matrix tri-factorizations for clustering, *KDD*, 126-135.

Ding,C., Li,T., and Jordan,M.I. (2010) Convex and semi-nonnegative matrix factorizations, *IEEE Transations on Pattern Analysis and Machine Intelligence*, **32(1)**, 45-55.

Ho, N.H. (2008) Nonnegative matrix factorization algorithms and applications. PhD Thesis, Catholic University of Louvain, Louvain-la-Neuve, Belgium.

Kim,H., and Park,H. (2007) Sparse non-negatice matrix factorization via alternating non-negativity-constrained least squares for microarray data analysis, *Bioinformatics*, **23(12)**, 1495-1502.

Lee,D.D., and Seung,S. (1999) Learning the parts of objects by non-negative matrix factorization, *Science*, **401**, 788-791.

Li,Y, and Ngom,A. (2010) Non-negative matrix and tensor factorization based classification of clinical microarray gene expression data, *BIBM*, 438-443.

Li,Y., Ngom,A. (2011) Non-negative least squares classifier, School of Computer Science, University of Windsor, available at `http://cs.uwindsor.ca/~li11112c/doc/nips2011.pdf`.

Li,Y. (2011) Non-negative matrix factorization classifier, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, submitted, available at `http://cs.uwindsor.ca/~li11112c/doc/tcbb.pdf`.

Madeira,S.C., and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a Survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **1(1)**, 24-45.