ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

Pareto front of bi-objective kernel-based nonnegative matrix factorization

Fei Zhu, Paul Honeine [*]

Institut Charles Delaunay (CNRS), Université de Technologie de Troyes, France

Abstract. The nonnegative matrix factorization (NMF) is a powerful data analysis and dimensionality reduction technique. So far, the NMF has been limited to a single-objective problem in either its linear or nonlinear kernel-based formulation. This paper presents a novel bi-objective NMF model based on kernel machines, where the decomposition is performed simultaneously in both input and feature spaces. The problem is solved employing the sum-weighted approach. Without loss of generality, we study the case of the Gaussian kernel, where the multiplicative update rules are derived and the Pareto front is approximated. The performance of the proposed method is demonstrated for unmixing hyperspectral images.

1 Introduction

The nonnegative matrix factorization (NMF) is a powerful data analysis and dimensionality reduction technique. It seeks to approximate a high-rank nonnegative input matrix by two nonnegative low-rank ones. A virtue of NMF is the ability to provide a parts-based representation for nonnegative input data, which facilitates a tractable physical interpretation. Suitable to the hyperspectral unmixing problem, the NMF jointly extracts the "pure" spectra called endmembers (recorded in the first low-rank matrix) and estimates the corresponding abundances of each endmember at each pixel (recorded in the second one).

Early NMF and its variants mainly consider a linear model. The objective function for minimization, namely the difference between the input matrix and the product of the estimated ones, is defined in an Euclidean space — the so-called *input space*. This common objective function with Frobenius norm [1] is often improved by additive regularization terms. Recent works have been extending the linear NMF model to the nonlinear scope, by exploiting the framework offered by the kernel machines. The kernel-based methods mainly map the data with some nonlinear function into a reproducing kernel Hilbert space — the so-called *feature space* —, where the existing linear techniques are performed on the transformed data. Unfortunately, most kernel-based NMF, *e.g.*, [2, 3], suffer from the pre-image problem [4], that is, the obtained bases lie in the feature space and a reverse mapping to the input space is difficult. In [5], we circumvent this problem by minimizing an objective function defined in the feature space.

So far, in either its linear conventional formulation or its nonlinear kernel-based formulation, as well as all of their variations, the NMF has been restricted to a single-objective optimization problem. In essence, the underlying assumption is that it is known in prior that certain model is most suitable to the data

[*] This work was supported by the French ANR, grand HYPANEMA: ANR-12BS03-0033.

under study. To obtain such prior information is not practical in most real-world applications. To this end, we propose to view the NMF as a multi-objective problem, in particular a bi-objective problem in this paper, where the objective functions defined in both input and feature spaces are taken into account. Instead of a single optimal decomposition, we seek a set of nondominated, Pareto, optimal solutions [6], by minimizing a weighted sum of objectives [7, 8].

2 A primer on the linear and kernel-based NMF

Given a nonnegative data matrix $\boldsymbol{X} \in \Re^{L \times T}$, the conventional NMF aims to approximate it by the product of two low-rank nonnegative matrices $\boldsymbol{E} \in \Re^{L \times N}$ and $\boldsymbol{A} \in \Re^{N \times T}$, *i.e.*, $\boldsymbol{X} \approx \boldsymbol{E}\boldsymbol{A}$. An equivalent vector-wise model is given by $\boldsymbol{x}_t$ for $t = 1, \ldots, T$, with $\boldsymbol{x}_t \approx \sum_{n=1}^{N} a_{nt}\, \boldsymbol{e}_n$, where each column of $\boldsymbol{X}$, namely $\boldsymbol{x}_t$, is represented as a linear combination of the columns of $\boldsymbol{E}$, denoted $\boldsymbol{e}_n$ for $n = 1, \ldots, N$, with the scalars a_{nt} for $n = 1, \ldots, N$ and $t = 1, \ldots, T$ being the entries of $\boldsymbol{A}$. The input space $\mathcal{X}$ is the space spanned by the vectors $\boldsymbol{x}_t$, as well as $\boldsymbol{e}_n$. To estimate the unknown matrices $\boldsymbol{E}$ and $\boldsymbol{A}$, one concentrates on the minimization of the Frobenius squared error norm $\frac{1}{2}\|\boldsymbol{X} - \boldsymbol{E}\boldsymbol{A}\|_F^2$, under nonnegativity constraints. In its vector-wise formulation, the objective function to minimize is

$$J_{\mathcal{X}}(\boldsymbol{E}, \boldsymbol{A}) = \frac{1}{2} \sum_{t=1}^{T} \|\boldsymbol{x}_t - \sum_{n=1}^{N} a_{nt}\, \boldsymbol{e}_n\|^2, \tag{1}$$

where the residual error is measured between each input vector $\boldsymbol{x}_t$ and its approximation $\sum_{n=1}^{N} a_{nt}\boldsymbol{e}_n$ in the input space $\mathcal{X}$. The objective function can be optimized by the two-block coordinate descent scheme, which alternates between the elements of $\boldsymbol{E}$ or of $\boldsymbol{A}$, by keeping the elements in the other matrix fixed.

We revisit also a kernel-based NMF presented in our recent works [5, 9]. Consider a nonlinear function $\Phi(\cdot)$ that maps the columns of matrices $\boldsymbol{X}$ and $\boldsymbol{E}$, from the input space $\mathcal{X}$ to some feature space $\mathcal{H}$. Its associated norm is denoted $\|\cdot\|_{\mathcal{H}}$, and the corresponding inner product in the feature space takes the form $\langle \Phi(\boldsymbol{x}_t), \Phi(\boldsymbol{x}_{t'}) \rangle_{\mathcal{H}}$, which can be evaluated using the so-called kernel function $\kappa(\boldsymbol{x}_t, \boldsymbol{x}_{t'})$ in kernel machines. Applying the vector-wise NMF model in the feature space, we get the following matrix factorization model $[\Phi(\boldsymbol{x}_1) \;\; \Phi(\boldsymbol{x}_2) \;\; \cdots \;\; \Phi(\boldsymbol{x}_T)] \approx [\Phi(\boldsymbol{e}_1) \;\; \Phi(\boldsymbol{e}_2) \;\; \cdots \;\; \Phi(\boldsymbol{e}_N)]\boldsymbol{A}$; or equivalently the vector-wise model $\Phi(\boldsymbol{x}_t) \approx \sum_{n=1}^{N} a_{nt}\, \Phi(\boldsymbol{e}_n)$, for all $t = 1, \ldots, T$. The optimization problem consists in minimizing the sum of the residual errors in the feature space $\mathcal{H}$, between each $\Phi(\boldsymbol{x}_t)$ and its approximation $\sum_{n=1}^{N} a_{nt}\, \Phi(\boldsymbol{e}_n)$, namely

$$J_{\mathcal{H}}(\boldsymbol{E}, \boldsymbol{A}) = \frac{1}{2} \sum_{t=1}^{T} \Big\|\Phi(\boldsymbol{x}_t) - \sum_{n=1}^{N} a_{nt}\, \Phi(\boldsymbol{e}_n)\Big\|_{\mathcal{H}}^2, \tag{2}$$

where the nonnegativity is imposed on all entries of $\boldsymbol{E}$ and $\boldsymbol{A}$. By analogy to the linear case, a two-block coordinate descent scheme can be investigated to solve this optimization problem.

3 The proposed bi-objective NMF

We propose to minimize the bi-objective function $[J_{\mathcal{X}}(\boldsymbol{E},\boldsymbol{A}) \quad J_{\mathcal{H}}(\boldsymbol{E},\boldsymbol{A})]$, under the nonnegativity of the matrices $\boldsymbol{E}$ and $\boldsymbol{A}$. The decision solution, of size $LN+NT$, corresponds to the entries in the unknown matrices $\boldsymbol{E}$ and $\boldsymbol{A}$. We adopt the well-known sum-weighted approach [7, 8], proposed in multiobjective optimization, to solve this bi-objective problem. This approach converts a multi-objective problem into a set of single-objective scalar problems (*i.e.*, suboptimization problems) by combining the multiple objectives. As proven in [7], the objective vector [1] corresponding to the optimal solution belongs to the convex part of multi-objective problem's Pareto front. Thus, by changing the weights among the objectives appropriately, the Pareto front of the original problem is approximated.

To this end, we consider the following suboptimization problem $\alpha J_{\mathcal{X}}(\boldsymbol{E},\boldsymbol{A})+(1-\alpha)J_{\mathcal{H}}(\boldsymbol{E},\boldsymbol{A})$, under the nonnegativity of the matrices $\boldsymbol{E}$ and $\boldsymbol{A}$. Here, the weight $\alpha\in[0,1]$ represents the relative importance between objectives $J_{\mathcal{X}}$ and $J_{\mathcal{H}}$. It is obvious that the model breaks down to the single-objective conventional NMF in (1) when $\alpha=1$, while the extreme case with $\alpha=0$ leads to the kernel-based NMF in (2). Let $J(\boldsymbol{E},\boldsymbol{A})=\alpha J_{\mathcal{X}}(\boldsymbol{E},\boldsymbol{A})+(1-\alpha)J_{\mathcal{H}}(\boldsymbol{E},\boldsymbol{A})$ be the aggregated objective function for some weight α. Substituting the expressions given in (1) and (2) for $J_{\mathcal{X}}$ and $J_{\mathcal{H}}$, it becomes

$$J=\frac{\alpha}{2}\sum_{t=1}^{T}\Big\|\boldsymbol{x}_t-\sum_{n=1}^{N}a_{nt}\,\boldsymbol{e}_n\Big\|^2+\frac{1-\alpha}{2}\sum_{t=1}^{T}\Big\|\Phi(\boldsymbol{x}_t)-\sum_{n=1}^{N}a_{nt}\,\Phi(\boldsymbol{e}_n)\Big\|_{\mathcal{H}}^2. \tag{3}$$

We apply the two-block coordinate descent scheme to solve the nonconvex problem. The derivative of (3) with respect to a_{nt} is

$$\nabla_{a_{nt}}J=\alpha\Big(-\boldsymbol{e}_n^\top\boldsymbol{x}_t+\sum_{m=1}^{N}a_{mt}\,\boldsymbol{e}_n^\top\boldsymbol{e}_m\Big)+(1-\alpha)\Big(-\kappa(\boldsymbol{e}_n,\boldsymbol{x}_t)+\sum_{m=1}^{N}a_{mt}\,\kappa(\boldsymbol{e}_n,\boldsymbol{e}_m)\Big), \tag{4}$$

while the gradient of (3) with respect to $\boldsymbol{e}_n$ satisfies

$$\begin{aligned}\nabla_{\boldsymbol{e}_n}J=&\ \alpha\sum_{t=1}^{T}a_{nt}\Big(-\boldsymbol{x}_t+\sum_{m=1}^{N}a_{mt}\boldsymbol{e}_m\Big)\\&+(1-\alpha)\sum_{t=1}^{T}a_{nt}\Big(-\nabla_{\boldsymbol{e}_n}\kappa(\boldsymbol{e}_n,\boldsymbol{x}_t)+\sum_{m=1}^{N}a_{mt}\,\nabla_{\boldsymbol{e}_n}\kappa(\boldsymbol{e}_n,\boldsymbol{e}_m)\Big).\end{aligned} \tag{5}$$

Here, $\nabla_{\boldsymbol{e}_n}\kappa(\boldsymbol{e}_n,\cdot)$ represents the gradient of the kernel with respect to its argu-

[1] The solution $(\boldsymbol{E}_1,\boldsymbol{A}_1)$ is said to *dominate* $(\boldsymbol{E}_2,\boldsymbol{A}_2)$ if and only if $J_{\mathcal{X}}(\boldsymbol{E}_1,\boldsymbol{A}_1)\leq J_{\mathcal{X}}(\boldsymbol{E}_2,\boldsymbol{A}_2)$ and $J_{\mathcal{H}}(\boldsymbol{E}_1,\boldsymbol{A}_1)\leq J_{\mathcal{H}}(\boldsymbol{E}_2,\boldsymbol{A}_2)$, where at least one inequality is strict. A given solution $(\boldsymbol{E}^*,\boldsymbol{A}^*)$ is a *Pareto optimal* if and only if it is not dominated by any other solution in the decision space. The set of the objective vectors corresponding to the Pareto optimal solutions forms the *Pareto front* in the objective space.

ment $\boldsymbol{e}_n$. A simple additive update rule takes the form

$$\begin{cases} a_{nt} & = a_{nt} = a_{nt} - \eta_{nt} \, \nabla_{a_{nt}} J; \\ \boldsymbol{e}_n & = \boldsymbol{e}_n - \eta_n \nabla_{\boldsymbol{e}_n} J, \end{cases} \tag{6}$$

where η_{nt} and η_n are the stepsize parameters which balance the rate of convergence with the accuracy of optimization and can be set differently depending on n and t. The additive update rule is easy to implement but the convergence can be slow and very sensitive to the stepsize value; also, a rectification function $a_{nt} = \max(a_{nt}, 0)$ is required along with the iteration to guarantee the nonnegativity of all a_{nt} and the entries in all $\boldsymbol{e}_n$.

Following the same spirit in Lee and Seung's paper [1], we provide the multiplicative update rules in the case of the Gaussian kernel in the second objective function $J_{\mathcal{H}}$. It is worth to emphasize that the multiplicative update rules for most valid kernels can be derived using a similar procedure. The Gaussian kernel is defined by $\kappa(\boldsymbol{z}_i, \boldsymbol{z}_j) = \exp(\frac{-1}{2\sigma^2}\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2)$ for any $\boldsymbol{z}_i, \boldsymbol{z}_j \in \mathcal{X}$, where σ denotes the tunable bandwidth parameter. Its gradient with respect to $\boldsymbol{e}_n$ is $\nabla_{\boldsymbol{e}_n} \kappa(\boldsymbol{e}_n, \boldsymbol{z}) = -\frac{1}{\sigma^2}\kappa(\boldsymbol{e}_n, \boldsymbol{z})(\boldsymbol{e}_n - \boldsymbol{z})$. By incorporating these expressions into (4) and (5), and choosing the stepsize parameter in (6) appropriately with the so-called gradient split method [2], we obtain the following multiplicative update rule

$$a_{nt} = a_{nt} \times \frac{\alpha \, \boldsymbol{e}_n^\top \boldsymbol{x}_t + (1-\alpha)\, \kappa(\boldsymbol{e}_n, \boldsymbol{x}_t)}{\alpha \displaystyle\sum_{m=1}^{N} a_{mt} \boldsymbol{e}_n^\top \boldsymbol{e}_m + (1-\alpha) \sum_{m=1}^{N} a_{mt}\, \kappa(\boldsymbol{e}_n, \boldsymbol{e}_m)}, \tag{7}$$

for a_{nt}, as well as

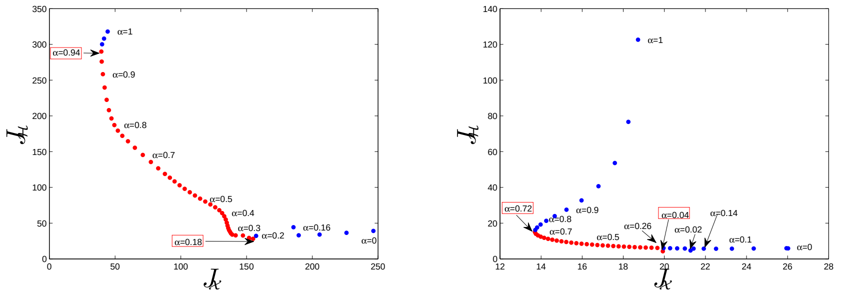
$$\boldsymbol{e}_n = \boldsymbol{e}_n \otimes \frac{\alpha\sigma^2 \sum_{t=1}^{T} a_{nt}\boldsymbol{x}_t + (1-\alpha)\sum_{t=1}^{T} a_{nt}\Big(\kappa(\boldsymbol{e}_n, \boldsymbol{x}_t)\boldsymbol{x}_t + \sum_{m=1}^{N} a_{mt}\kappa(\boldsymbol{e}_n, \boldsymbol{e}_m)\boldsymbol{e}_n\Big)}{\alpha\sigma^2 \sum_{t=1}^{T} a_{nt} \sum_{m=1}^{N} a_{mt}\boldsymbol{e}_m + (1-\alpha)\sum_{t=1}^{T} a_{nt}\Big(\kappa(\boldsymbol{e}_n, \boldsymbol{x}_t)\boldsymbol{e}_n + \sum_{m=1}^{N} a_{mt}\kappa(\boldsymbol{e}_n, \boldsymbol{e}_m)\boldsymbol{e}_m\Big)}, \tag{8}$$

for $\boldsymbol{e}_n$, where the division and multiplication are element-wise.

4 Pareto front for unmixing hyperspectral images

The efficiency of the proposed bi-objective NMF is demonstrated for unmixing two sub-images (of size 50 × 50 pixels) taken respectively from the Urban and Cuprte image [10]. Experiments are conducted employing the weight set $\alpha \in \{0, 0.02, ..., 0.98, 1\}$. For each α from the set, multiplicative update rules given in (7)-(8) are applied, with the maximum iteration number $n_{\max} = 300$. We fix the bandwidth in the Gaussian kernel as $\sigma = 3.0$ for the Urban image, and $\sigma = 2.5$ for the Cuprite image. To approximate the Pareto front with a discrete set of points, we operate as follows: For each value of the weight α, we obtain a solution (endmember and abundance matrices) with the multiplicative update

[2] The gradient split method decomposes the expression of the gradient into the subtraction of two nonnegative terms, *i.e.*, $\nabla_{\boldsymbol{e}_n} J = P - Q$, where P and Q have nonnegative entries.



(a) Urban image

(b) Cuprite image

Fig. 1: Illustration of the approximated Pareto front in the objective space. The objective vectors of the non-dominated solutions (42 for the Urban image, 28 for the Cuprite image), marked in red, approximate a part of the Pareto front; the objective vectors of the dominated solutions (9 for the Urban image, 23 for the Cuprite image) are marked in blue.

rules (7) and (8); by evaluating the objective functions $J_{\mathcal{X}}$ and $J_{\mathcal{H}}$ at this solution, we get a single point in the objective space. The approximated Pareto front for the Urban and the Cuprite images are shown in Fig. 1(a) and Fig. 1(b), respectively. We observe the following:

1) For both images under study, solutions generated with $\alpha = 1$ and $\alpha = 0$ are dominated, since all the solutions on the Pareto front outperform them, with respect to both objectives. This reveals that neither the conventional linear NMF nor the nonlinear Gaussian kernel NMF best fits the studied images. On the contrary, the Pareto optimal solutions, which result in the points on the Pareto front, provide a set of feasible and nondominated decompositions.
2) Theoretically, the minimizer of the suboptimization problem is a Pareto optimal for the original multiobjective problem. In practice, we obtain 9 and 23 (out of 51) dominated solutions for the Urban and the Cuprite images, respectively. This phenomenon, however, is not surprising and could be caused by the failure of the solver in finding a global minimum [6].
3) An even distribution of weight α between [0, 1] do not lead to an even spread of the solutions on the approximated Pareto front. Moreover, the nonconvex part of the Pareto front cannot be attained using any weight. It is exactly the case in Fig. 1(b); in Fig. 1(a), a trivial nonconvex part between $\alpha = 0.3$ and $\alpha = 0.5$ on the approximated Pareto front is probably resulted from the nonoptimal solution of the suboptimization problem. These are two main drawbacks of the sum-weighted method.

Due to the page limitation, we omit the unmixing results comparison with the state-of-the-art methods. Relative results in terms of the reconstruction errors in both input and feature spaces, as well as the resulting endmembers with their corresponding abundance maps, can be found in the extended paper [11]. The obtained approximation of Pareto front is of high value. On one hand, it provides a set of Pareto optimal solutions for the user, instead of a single decomposition. On the other hand, an insight of the tradeoff between objectives $J_{\mathcal{X}}$ and $J_{\mathcal{H}}$ reveals the underlying linearity/nonlinearity of the data under study.

5 Conclusion

This paper presented a novel bi-objective nonnegative matrix factorization by exploiting the kernel machines, where the decomposition was performed simultaneously in input and feature space. The multiplicative update rules were derived. The performance of the method was demonstrated for unmixing well-known hyperspectral images. The resulting Pareto front was analyzed. As for future work, we are extending this approach to include other NMF objective functions, defined in the input or the feature space. Considering simultaneously several kernels, and as a consequence several feature spaces, is also under investigation, following the wide study in multiple kernel learning [12].

References

[1] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[2] D. Zhang, Z. Zhou, and S. Chen. Non-negative matrix factorization on kernels. In *Lecture Notes in Computer Science*, volume 4099, pages 404–412. Springer, 2006.

[3] C. Ding, T. Li, and M. I. Jordan. Convex and Semi-Nonnegative Matrix Factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55, Nov. 2010.

[4] P. Honeine and C. Richard. Preimage problem in kernel-based machine learning. *IEEE Signal Processing Magazine*, 28(2):77–88, 2011.

[5] F. Zhu, P. Honeine, and M. Kallas. Kernel non-negative matrix factorization without the pre-image problem. In *IEEE workshop on Machine Learning for Signal Processing (MLSP)*, Reims, France, Sep. 2014.

[6] K. Miettinen. Introduction to multiobjective optimization: Noninteractive approaches. In *Multiobjective Optimization*, pages 1–26. Springer, 2008.

[7] I. Das and J.E. Dennis. A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems. *Structural optimization*, 14(1):63–69, 1997.

[8] J. Ryu, S. Kim, and H. Wan. Pareto front approximation with adaptive weighted sum method in multiobjective simulation optimization. In *Simulation Conference (WSC), Proceedings of the 2009 Winter*, pages 623–633, Dec. 2009.

[9] F. Zhu, P. Honeine, and M. Kallas. Kernel nonnegative matrix factorization without the curse of the pre-image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (submitted in 2014) 2015 preprint.

[10] J. Chen, C. Richard, and P. Honeine. Nonlinear unmixing of hyperspectral data based on a linear-mixture/nonlinear-fluctuation model. *IEEE Transactions on Signal Processing*, 61(2):480–492, Jan. 2013.

[11] F. Zhu and P. Honeine. Bi-objective nonnegative matrix factorization: Linear versus kernel-based models. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–11, (submitted in 2014) 2015 preprint.

[12] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, Jul. 2011.