ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

ONP-MF: An Orthogonal Nonnegative Matrix Factorization Algorithm with Application to Clustering

Filippo Pompili[1], Nicolas Gillis[2], P.-A. Absil[2], and François Glineur[2,3]

1- University of Perugia, Department of Electronic and Information Engineering
Via G. Duranti 93, I-06125 Perugia, Italy

2- Université catholique de Louvain, ICTEAM Institute
B-1348 Louvain-la-Neuve, Belgium

3- Université catholique de Louvain, CORE
Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium

Abstract. Given a nonnegative matrix M, the orthogonal nonnegative matrix factorization (ONMF) problem consists in finding a nonnegative matrix U and an orthogonal nonnegative matrix V such that the product UV is as close as possible to M. The importance of ONMF comes from its tight connection with data clustering. In this paper, we propose a new ONMF method, called ONP-MF, and we show that it performs in average better than other ONMF algorithms in terms of accuracy on several datasets in text clustering and hyperspectral unmixing.

1 Introduction

We consider the orthogonal nonnegative matrix factorization (ONMF) problem, which can be formulated as follows. Given an m-by-n nonnegative matrix M and a factorization rank k (with $k < n$), solve

$$\min_{U \in \mathbb{R}^{m \times k}, V \in \mathbb{R}^{k \times n}} ||M - UV||_F^2 \tag{1a}$$

$$\text{subject to} \quad U \geq 0,\ V \geq 0, \tag{1b}$$

$$VV^{\mathrm{T}} = \mathrm{I}_k, \tag{1c}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, (1b) means that the entries of matrices U and V are nonnegative, and I_k stands for the $k \times k$ identity matrix. The ONMF problem (1) can be viewed as the well-known nonnegative matrix factorization (NMF) problem, (1a)-(1b), with an additional orthogonality constraint, (1c), that considerably modifies the nature of the problem. In particular, it is readily seen that constraints (1b) and (1c) imply that V has at most one nonzero entry in each column; we let i_j denote the index of the nonzero entry (if any) in column j of V. Therefore, any solution (U^*, V^*) of (1) has the following property: for $j = 1, \ldots, n$, index i_j is such that column i_j of U^* achieves the smallest angle with column j of data matrix M. Hence it is clear that the ONMF problem relates to data clustering and, indeed, empirical evidence suggests that the additional orthogonality constraint (1c) can improve clustering performance compared to standard NMF or k-means [1, 2].

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Current approaches to ONMF problems are based on suitable modifications of the algorithms developed for the original NMF problem. They enforce nonnegativity of the iterates at each step, and strive to attain orthogonality at the limit (but never attain exactly orthogonal solutions). This can be done using a proper penalization term [3], a projection matrix formulation [2] or by choosing a suitable search direction [1]. Note that, for a given data matrix M, different methods may converge to different pairs (U, V), where the objective function (1a) may take different values. Furthermore, under random initialization, which is used by most NMF algorithms [4], two runs of the same method may yield different results. This situation is due to the multimodal nature of the ONMF problem (1)—it may have multiple local minima—along with the inability of practical methods to guarantee more than convergence to local, possibly nonglobal, minimizers. Hence, ONMF methods not only differ in their computational cost, but also in the quality of the clustering encoded in the returned pair (U, V) for a given problem.

In this paper, we propose a new ONMF method, referred to as orthogonal nonnegatively penalized matrix factorization (ONP-MF), that relies on a strategy reversal: instead of enforcing nonnegativity of the iterates at each step and striving to attain orthogonality at the limit, ONP-MF enforces orthogonality of its iterates while obtaining nonnegativity at the limit. A resulting advantage of ONP-MF is that rows of factor V can be initialized directly with the right singular vectors of M, whereas the other methods require a prior alteration of the singular vectors that makes them nonnegative [4]. We show that, on some clustering problems, the new algorithm performs in average better than standard ONMF algorithms in terms of clustering quality.

The proposed ONP-MF is introduced in Section 2, numerical experiments are presented in Section 3, and conclusions are drawn in Section 4. An early account of ONP-MF can be found in the technical report [5], where additional numerical experiments are presented.

2 The ONP-MF algorithm

For an optimization problem such as (1), a standard bound-constrained Lagrangian approach [6, Algorithm 17.4] would typically only incorporate the orthogonality constraint (1c) in the augmented Lagrangian, while nonnegativity constraints (1b) would be enforced explicitly. This yields an algorithm in the spirit of the current approaches mentioned in the introduction. However, problem (1) is special in that the equality constraints (1c) take a gentle form. Indeed, the feasible set for (1c), $\mathrm{St}(k, n) := \{V \in \mathbb{R}^{k\times n} : VV^{\mathrm{T}} = \mathrm{I}_k\}$, is a well-known manifold termed *Stiefel manifold*; see, e.g., [7, 8]. In particular, as we will see, enforcing the orthogonality constraint (1c) can be done at a low computational cost. This prompts us to take a reverse approach and enforce the orthogonality constraint (1c) while incorporating the nonnegativity constraints on V in the augmented Lagrangian. The nonnegativity constraints on U remain explicitly enforced. We now work out the optimization scheme in more details.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Consider the following augmented Lagrangian, defined for a matrix of Lagrange multipliers $\Lambda \in \mathbb{R}_+^{k\times n}$ associated to the nonnegativity constraints on V:

$$L_\rho(U,V,\Lambda) = \frac{1}{2}||M-UV||_F^2 + \langle \Lambda, -V\rangle + \frac{\rho}{2}||\min(V,0)||_F^2, \tag{2}$$

where ρ is the (positive) quadratic penalty parameter. Observe that, regardless of the value of ρ, the solutions (U,V) of ONMF (1) are the solutions of

$$\min_{U\geq 0, VV^{\mathrm{T}}=\mathrm{I}_k} \max_{\Lambda\geq 0} L_\rho(U,V,\Lambda).$$

We propose a simple alternating scheme to update variables U, V, Λ while explicitly enforcing $U \geq 0$ and $VV^{\mathrm{T}} = \mathrm{I}_k$:

1. For V and Λ fixed, the optimal U can be computed by solving a nonnegative least squares problem, i.e., by letting $U \leftarrow \operatorname{argmin}_{X\in\mathbb{R}_+^{m\times k}} \|M - XV\|_F^2$. We use the efficient active-set method proposed in [9].[1]

2. For U and Λ fixed, we update matrix V by means of a projected gradient step. Computing the projection of a matrix $\hat{V}$ onto the feasible set $\mathrm{St}(k,n)$ of orthogonal matrices amounts to solving the following problem: $\mathrm{Proj}_{\mathrm{St}}(\hat{V}) = \operatorname{argmin}_X ||\hat{V} - X||_F^2$ s.t. $XX^{\mathrm{T}} = \mathrm{I}_k$. When $\hat{V}$ has full (row) rank, the solution is unique and given by $(\hat{V}\hat{V}^{\mathrm{T}})^{-1/2}\hat{V}$, which can be recognized as the orthonormal factor of a reverse polar decomposition of $\hat{V}$; see, e.g., [10] or [11, §3.3]. Our projected gradient scheme then reads:

$$V \leftarrow \mathrm{Proj}_{\mathrm{St}}\Big(V - \beta\nabla_V L_\rho(U,V,\Lambda)\Big), \tag{3}$$

   where the step length β is chosen with a backtracking line search as in [12].

3. Finally, Lagrange multipliers are updated in order to penalize the negative entries of V: $\Lambda \leftarrow \max(0, \Lambda - \alpha V)$. As $-V$ is the gradient of function $\Lambda \mapsto L_\rho(U,V,\Lambda)$, this update is a (projected) gradient step with step length α. We choose a predefined step length sequence $\alpha = \alpha_0/t$, where t is the iteration count and $\alpha_0 > 0$ is a constant parameter, that satisfies the usual "square summable but not summable" condition of online gradient methods [13, (5.1)].

To initialize the algorithm, we set Λ to zero and choose for the rows of V the first k right singular vectors of the data matrix M, which can be obtained with a singular value decomposition (SVD). The quadratic penalty parameter ρ is initially fixed to a given small value ρ_0 and then increased after each iteration using $\rho \leftarrow C\rho$ for some $C > 1$. The parameters of ONP-MF are chosen as follows: $\alpha_0 = 100$, $\rho_0 = 0.01$ and $C = 1.01$ for *all* datasets. Our alternating procedure will be referred to as orthogonal nonnegatively penalized matrix factorization (ONP-MF).

[1] Software: http://www.cc.gatech.edu/~hpark/nmfsoftware.php.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

3 Numerical Experiments

In this section, we report some preliminary numerical experiments that illustrate the clustering ability of ONP-MF in comparison with various state-of-the-art algorithms (a more detailed comparison is under investigation, and not possible here due to the space limitation). First, we compare ONP-MF with two recently proposed methods for ONMF: CHNMF from Choi [1] and PNMF from Yang and Oja [2] (Euclidean variant). Next, since our ONP-MF is initialized deterministically using an SVD whereas the other ONMF algorithms (CHNMF and PNMF) are initialized with randomly generated factors, we propose a fairer comparison by also endowing CHNMF and PNMF with an SVD-based initialization [4] (SVD cannot be used directly because its factors are not necessarily nonnegative). We call the resulting algorithms CH(SVD) and P(SVD), respectively. Finally, we also report the results from two standard EM clustering algorithms, namely k-means and spherical k-means (SKM) (see, e.g., [14]). ONMF algorithms are run until an algorithm-specific stopping condition is met, or a maximum number of iterations is reached. All three ONMF algorithms have roughly the same computational cost, linear in m and n (that is, $\mathcal{O}(mn)$ operations per iteration), although it appears that ONP-MF is slightly slower in practice; see [5] for more details on stopping conditions, scalability and learning times. However, as we will see below, ONP-MF has the advantage that a single SVD initialization is very effective. For all algorithms but ONP-MF, 30 runs with randomly generated initializations were executed for each dataset.

3.1 Text clustering

We selected nine well-known preprocessed document databases described in [15]. Each dataset is represented by a term-by-document matrix of varying charac-

Table 1: Average accuracy obtained by the different algorithms (best in bold).

Dataset	k-means	SKM	CHNMF	CH(SVD)	PNMF	P(SVD)	ONP-MF
classic	**0.627**	0.575	0.543	0.559	0.534	0.548	0.538
ohscal	0.286	**0.435**	0.333	0.339	0.343	0.353	0.340
hitech	0.319	**0.492**	0.411	0.414	0.443	0.471	0.470
reviews	0.441	**0.676**	0.523	0.494	0.533	0.503	0.510
sports	0.395	0.452	0.431	0.491	0.443	0.410	**0.500**
la1	0.359	0.475	0.519	0.444	0.568	**0.660**	0.658
la2	0.330	0.483	0.435	0.422	0.472	0.510	**0.528**
k1b	0.681	0.647	0.750	0.606	0.755	0.782	**0.790**

teristics. As a performance indicator, we use the *accuracy*: given a clustering $\{\pi_i\}_{i=1}^k$ and the true classes $\{L_i\}_{i=1}^k$ of the n elements of the dataset, we have

$$\text{Accuracy} = \max_{P\in[1,2,\ldots,k]} \frac{1}{n}\left(\sum_{i=1}^{k} |\pi_i \cap L_{P(i)}|\right) \in [0,1],$$

where $[1,2,\ldots,k]$ is the set of permutations of $\{1,2,\ldots,k\}$. Accuracies obtained by the tested algorithms are reported in Table 1 (for algorithms with random

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

initialization, average accuracy is displayed). ONP-MF achieved the best performance on three out of eight datasets (best method, tied with SKM), and its performance was within 5% of the best on five out of eight datasets (best method according to this criterion, the second best being P(SVD) with four out of eight). In particular, ONP-MF performs in average better than all other ONMF algorithms on these datasets.

3.2 Hyperspectral Unmixing

A hyperspectral image is a set of images of the same object or scene taken at different wavelengths. Each image is acquired by measuring the reflectance (i.e., the fraction of incident electromagnetic power reflected) of each individual pixel at a given wavelength. Our goal is to recognize the different materials appearing in the image, such as grass, metal, etc. More precisely, we want to cluster the columns of the wavelength-by-pixel reflectance matrix so that each cluster (a set of pixels) corresponds to a particular material. In this paper, we report results for a dataset on which superior performance of ONP-MF is particularly obvious; more results can be found in the technical report [5]. The dataset is a synthetic dataset from [16], see Figure 1 (top row), in clean conditions (i.e., without noise or blur). It represents the Hubble telescope and is made up of eight different materials, each having a specific spectral signature. Recall that each randomly initialized algorithm is run 30 times, and we display the best solution obtained (w.r.t. the error). Figure 1 displays the clustering obtained by the different algorithms. Unlike the other methods, ONP-MF is able to successfully recover all eight materials without any visible mixing (using a single initialization).

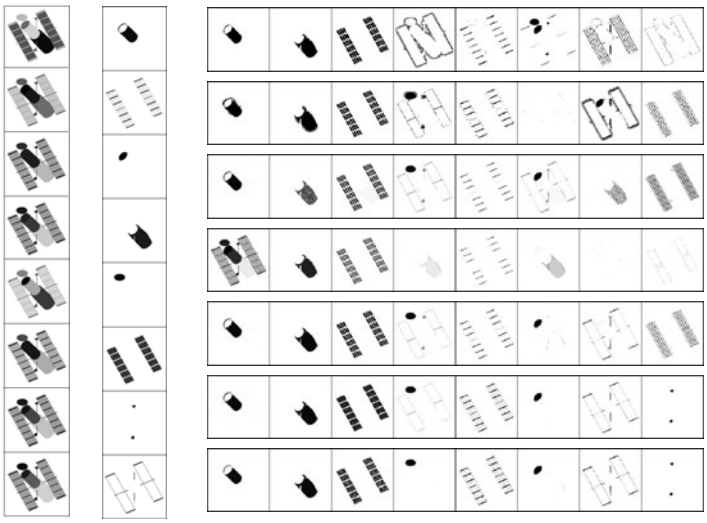

Fig. 1: Hubble dataset. On the left: sample images and true materials; from top to bottom: k-means, SKM, CHNMF, CH(SVD), PNMF, P(SVD), ONP-MF.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

4 Conclusion

We have proposed ONP-MF, an ONMF method that, in contrast with existing methods, enforces the orthogonality condition (1c) at each iteration while obtaining nonnegativity of the factors at the limit. In spite of the simple implementation of the concept, based on an alternating scheme applied to an augmented Lagrangian, the new method is observed to perform in average better than standard ONMF algorithms in terms of solution quality on several datasets. Since initialization is known to be an important component in the design of successful NMF methods [4], we believe that initializing the V factor with the unaltered right singular vectors of the data matrix, which is allowed by the workings of ONP-MF but impossible with other ONMF methods, plays an instrumental role in the clustering performance of ONP-MF observed in numerical experiments.

References

[1] S. Choi. Algorithms for orthogonal nonnegative matrix factorization. In *Proc. of the Int. Joint Conf. on Neural Networks*, pages 1828–1832, 2008.

[2] Z. Yang and E. Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 21:734–749, May 2010.

[3] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proc. of the Twelfth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 126–135, 2006.

[4] C. Boutsidis and E. Gallopoulos. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.

[5] F. Pompili, N. Gillis, P.-A. Absil, and F. Glineur. Two algorithms for orthogonal nonnegative matrix factorization with application to clustering. arXiv:1201.0901, 2012.

[6] J. Nocedal and S.J. Wright. *Numerical Optimization, Second Edition*. Springer, New York, 2006.

[7] A. Edelman, T. A. Arias, and S.T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1998.

[8] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[9] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM J. Matrix Anal. Appl.*, 30(2):713–730, 2008.

[10] R.A. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[11] P.-A. Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM J. Optim.*, 22(1):135–158, 2012.

[12] C.-J. Lin. Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural Computation*, 19:2756–2779, 2007. MIT press.

[13] L. Bottou. Online algorithms and stochastic approximations. In D. Saad, editor, *Online Learning and Neural Networks*. Cambridge University Press, 1998.

[14] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra. Generative model-based clustering of directional data. In *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-03)*, pages 19–28. ACM Press, 2003.

[15] S. Zhong and J. Ghosh. Generative model-based document clustering: a comparative study. *Knowledge and Information Systems*, 8(3):374–384, 2005.

[16] V.P. Pauca, J. Piper, and R.J. Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and its Applications*, 406(1):29–47, 2006.