ON AFFINE NON-NEGATIVE MATRIX FACTORIZATION

Hans Laurberg[1] *and Lars Kai Hansen*[2]

[1] Department of of Electronic Systems, Aalborg University,
Fredrik Bajers Vej 7A-3, DK-9220 Aalborg, Denmark,
[2] Informatics and Mathematical Modelling, Technical University
of Denmark B321, DK-4000 Lyngby, Denmark,
emails: {hla}@kom.aau.dk,{lkh}@imm.dtu.dk

ABSTRACT

We generalize the non-negative matrix factorization (NMF) generative model to incorporate an explicit offset. Multiplicative estimation algorithms are provided for the resulting sparse affine NMF model. We show that the affine model has improved uniqueness properties and leads to more accurate identification of mixing and sources.

Index Terms— Non-negative matrix factorization, NMF, BSS, Sparse NMF

1. INTRODUCTION

Non-negative matrix factorization (NMF) has become a popular tool for data analysis. An often stated reason for NMF is that it leads to 'parts based' representations, hence, facilitates data analytic interpretation. However, uniqueness is important for the parts based representations to be meaningful. The NMF generative model is based on linear mixing of positive sources by positive coefficients. The positive sources may have offsets which can lead to non-uniqueness, we therefore here propose a model based on *affine mixing*, i.e., mixing with an offset. The NMF learning algorithm is straightforwardly generalized to handle the augmented model. We show that the affine model indeed has improved uniqueness properties and thus leads to more accurate identification of mixing and sources.

NMF algorithms are used to factorize a nonnegative matrix $V \in \mathbb{R}^{N\times M}$ in two nonnegative matrices $W \in \mathbb{R}^{N\times D}$ and $H \in \mathbb{R}^{D\times M}$

$$V \approx R = WH;\;\; V_{i,j} \approx R_{i,j} = \sum_{d=1}^{D} W_{i,d}H_{d,j} \qquad (1)$$

Following the seminal papers by Lee and Seung [1, 2], a least squares or a Kullback-Leibler inspired cost are used. Our observations in this paper can be applied to both. For simplicity we will concentrate on the Euclidian cost in the following,

$$E(W,H) = \|V - WH\|_F^2\,, \qquad (2)$$

where $\|\cdot\|_F$ is the Frobenius norm. Lee and Seung[2] have shown that the following update rule will decrease $E(W,H)$

$$H \leftarrow H \otimes \frac{W^TV}{W^TR} \qquad (3)$$

$$W \leftarrow W \otimes \frac{VH^T}{RH^T}\;\;, \qquad (4)$$

where $\otimes$ and $\frac{(\cdot)}{(\cdot)}$ are element wise multiplication and division. This update rule is used as a reference and is shown in panel (B) of figures 1, 3, 4, 5 and 6.

2. SPARSE NMF

Hojer [3] introduced sparse NMF and Eggert[4] proposed the following cost function where only the normalized version of W has impact on the cost

$$E(W,H) = \frac{1}{2}\left\|V - \overline{W}H\right\|_F^2 + \lambda \mathbf{1}^T H \mathbf{1} \qquad (5)$$

$$\overline{W}_n = \frac{W_n}{\|W_n\|}\;\;, n \in \{1,\ldots,N\} \qquad (6)$$

where W_n is the n'th column vectorin W and $\mathbf{1}$ is a column vector where all elements are one. The length of $\mathbf{1}$ can be deduced by the context. The scalar λ is a positive parameter that controls the tradeoff between sparseness of H and approximation of V by the product of W, H. Eggert[4] argues for using the following multiplicative update

$$H \leftarrow H \otimes \frac{\overline{W}^TV}{\overline{W}^TR + \lambda} \qquad (7)$$

$$W_n \leftarrow \overline{W}_n \otimes \frac{\sum_{m=1}^{M} H_{m,n}(V_n + \overline{W}_n(R_m)^T\overline{W}_n)}{\sum_{m=1}^{M} H_{m,n}(R_n + \overline{W}_n(V_m)^T\overline{W}_n)} \qquad (8)$$

These update rules are used in panel (C) of figures 1, 3, 4, 5 and 6.

This research was supported by the Intelligent Sound project, Danish Technical Research Council grant no. 26–02–0092.

The normalization of W and the sparse nature of H critically constrains the solution and can improve uniqueness and lead to more accurate estimates. However, the constraints may not be consistent with the form of the mixing process and the statistics of the source signals H. In particular offsets in one or more rows of V will counteract the sparse model. If the generative model incorporates additive noise it is not clear that simple subtraction of the minimal value of each row in V will lead to a correct recovery of the generating W, H. If the noise is, e.g., Gaussian, V can be negative in the native representation, hence, one cannot estimate the 'true' offset.

2.1. Affine Sparse NMF

The above sparse NMF methods do not handle offsets, however, it is incorporated as follows with $W_0 \in \mathbb{R}^{N\times 1}$

$$V \approx R = WH + W_0\mathbf{1}^T. \tag{9}$$

Using this augmented signal model the sparse cost function in Equation 5 becomes

$$E(W,H,W_0) = \frac{1}{2}\left\|V - \overline{W}H - W_0\mathbf{1}^T\right\|_F^2 + \lambda\mathbf{1}^T H\mathbf{1} \tag{10}$$

Following Eggert[4] the update rule for W and H remains as given in Equation 7 and 8 using the new definition of R and the update rule for W_0 (that in not normalized) is the standard NMF update rule in Equation 4

$$W_0 \leftarrow W_0 \otimes \frac{\mathbf{1}^T V}{\mathbf{1}^T R} \tag{11}$$

The affine sparse NMF results are shown in panel (D) of figures 1, 3, 4, 5 and 6.

3. RESULTS

How does the augmented sparse affine NMF model data? To answer this question we first visualize synthetic data as generated by the proposed model, and we show existing methods fail to reconstruct the correct parameters of the generative model. We then go on to show that two commonly used data sets have the characteristics of the proposed model and that the proposed algorithm performs better than the existing algorithms on the data. In order to get a 'fair' comparison the standard NMF and sparse NMF both have one column more than the sparse affine NMF method. This ensures that the maximum rank of R is the same for all methods.

Simulated Data. In Figure 1 there are $M = 2000$ elements in V. The data is generated as in Equation 9. The elements of R are exponentially distributed. The true W vectors and the column vectors of V are shown in Figure 1 panel (A). Figure 1 (B–D) shows the three different algorithms estimate of W. The standard NMF (B) finds W such that the data is in the positive span of W. The W estimated by the sparse

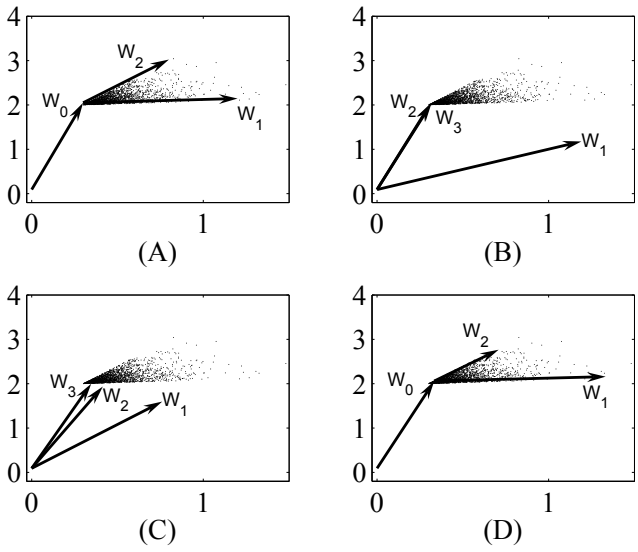

Fig. 1. Simulated data where $V \in \mathbb{R}^{2\times 2000}$ is generated according to Equation 9. Each column of V is plotted as a dot. In (A) the generating W and W_0 are shown. In (B) and (C) the standard NMF and sparse NMF each find three vectors that can describe the data. Both algorithms find one vector that is a linear combination of the true W_0 and W_1 and finds two vectors that are very close to the true W_0. In (D) the 'Affine sparse NMF' method correctly estimates the structure of the W matrix.

NMF algorithm (C) also spans data but the column vectors of W point more directly towards data. Although these methods estimated W can reproduce V, they do not find the correct structure (W). The proposed method (D) finds a W that is close to the true W.

A quantitative evaluation of the different algorithms' estimate is presented in Figure 2. Data is generated as in Equation 9 where the elements of W and W_0 are uniform i.i.d. The elements of H are first generated as exponential i.i.d. samples and then each column is normalized to unit sum. In this way the elements in H describe how much each column vector of W contribute towards V. In all simulations $N = 100, D = 10$. We have run the simulation with different amounts of data examples (column in V) M. In the evaluation V is analysed as 11(=D +1) outer product $\sum_{d=0}^{D} V^{(d)} = V$, where $V_{i,j}^{(d)} = W_{i,d}H_{d,j}$. The error in the figure is the relative least squares error of the $V^{(d)}$ estimate for each data set size

$$\frac{\sum_{d=0}^{D}\left\|V^{(d)} - R^{(d)}\right\|_F^2}{\sum_{d=0}^{D}\left\|V^{(d)}\right\|_F^2} \tag{12}$$

For completeness we have included in the performance evaluation a modification of the standard method in which data is first subtracted with constant offsets to achieve zero minimum value in each of the N variables of V. The simulation

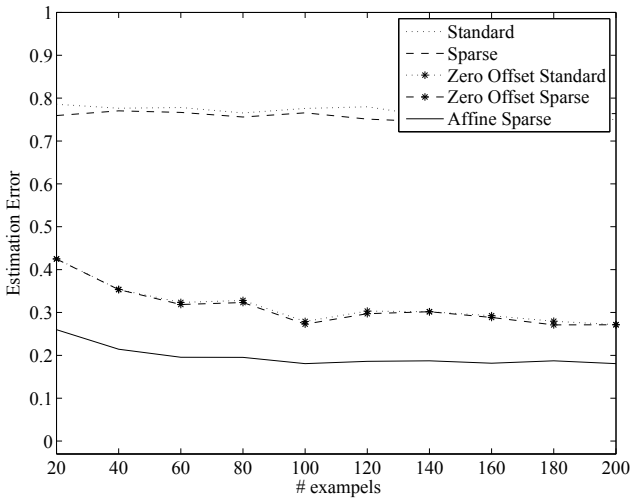

Fig. 2. The variation of the relative least squares error of the NMF reconstruction of W. The error is plotted as a function of the amount of data (M). The simulated data was generated using $D = 10$ components and an off set. The 'zero offset' methods are based on the simple heuristic that data is first preprocessed to have minimum value zero in each row.

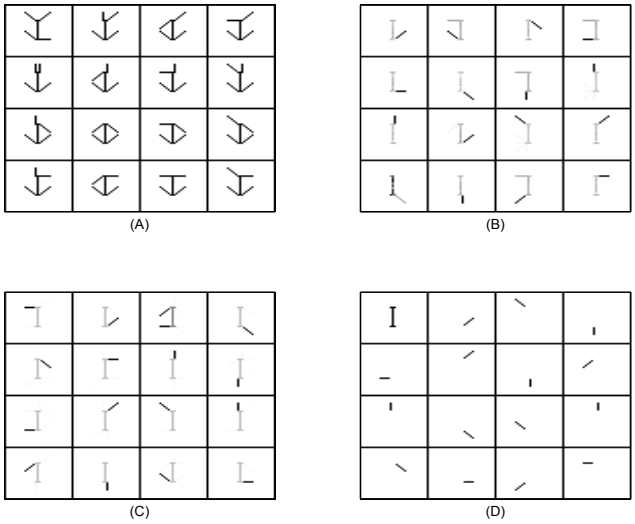

Fig. 3. Subset of A: The Swimmer database B: Basis pictures using standard NMF. C: Basis pictures using sparse NMF. D: Basis pictures using sparse affine NMF.

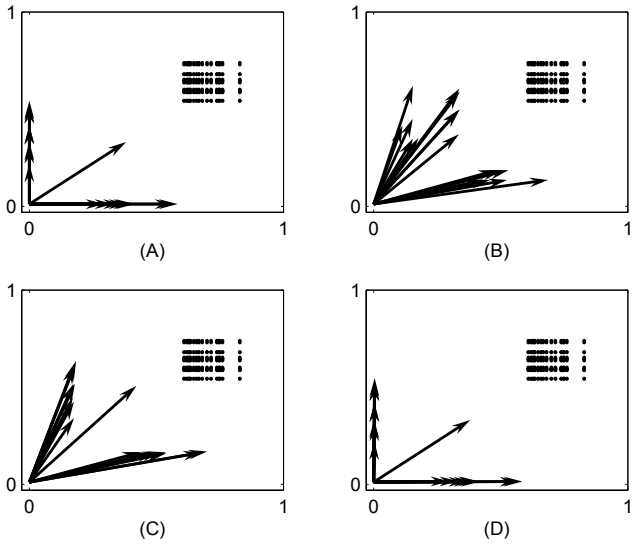

Fig. 4. A two dimensional subspace of the column vectors in V (dots) and W (vectors) are shown for the Swimmer database. The 'x-axis' is a picture which is zero in the upper part and uniform random values in the lower part. The 'y-axis' is constructed the same way but with the zeros in the lower part.

shows that the standard NMF and the sparse NMFs do not find the true W and H. The constant offset subtraction improves the performans but is outperformed by the sparse affine NMF succeeds. Notis that the two latter methods is favoured by knowing that $(H^T)_0 = \mathbf{1}$.

The Swimmer Database. The "Swimmer Database" was introduced by Donoho and Stodden [5] to discuss the uniqueness issues we have adressed in this presentation. The point was that even if NMF can represent V it may not necessarily find the right W. The database consist of 256 (32×32 pixel) black-and-white pictures of a 'stick-man' with 4 limbs that can be in one of 4 positions. All pictures have a 'torso' that represent an offset as discussed in this paper. The pictures in the dataset can be constructed by 17 ($= 4 \times 4 + 1$) non-overlapping basis pictures. In Figure 3 (A) examples from the database are shown. The algorithms described in section 2 are tested on the data set and a subset of the 17 basis pictures are shown in Figure 3(B–D). Only the proposed method is able to find the 17 non-overlapping basis pictures, the standard NMF and Sparse NMF all let the torso be a part of all basis pictures. The Swimmer simulation is further analyzed in Figure 4. The 1024 ($= 32 \times 32$) dimensional column vectors in V and W are mapped onto a two dimensional subspace to show that the structure of the swimmer database is in fact equivalent to that of Figure 1. In the plot it is seen that only the affine sparse NMF finds the true basis vectors.

Business Card Data Set. Our final example is based on a set of business card images of faculty of Aalborg University's Department of Electronic Systems. The photographer has manually centered and scaled the pictures. The pictures are scaled to 30×40 pixel and the color map is chosen such that white is zero and black is maximum. An 'AAU watermark' logo has been added to all pictures in the database. A subset of the pictures are shown in Figure 5(A) and a subset of the 25 basis pictures estimated by the three algorithms is

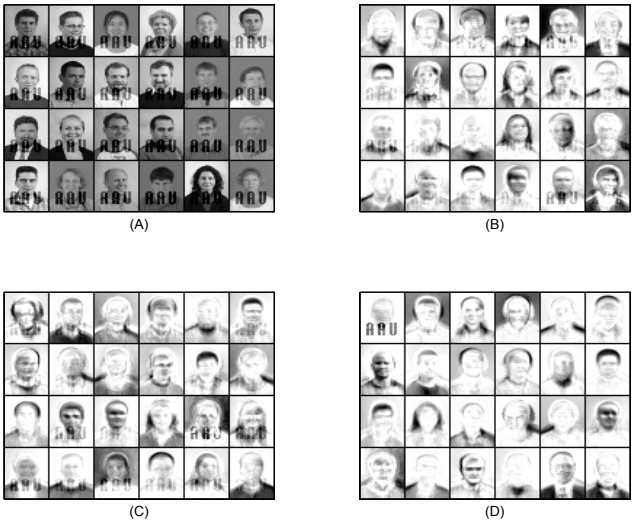

Fig. 5. (A): Subset of the Picture database with 197 pictures (B – D): A subset of the basis pictures using standard NMF, sparse NMF and sparse affine NMF. The standard NMF makes very noisy basis pictures. The sparse NMF produce basis pictures where the 'AAU watermark' is visible in around 50% of the pictures, and in addition a lot of the pictures do not represent a single part of the picture. The sparse affine NMF has only one picture with the watermark (W_0) and most pictures represent only one part of the picture.

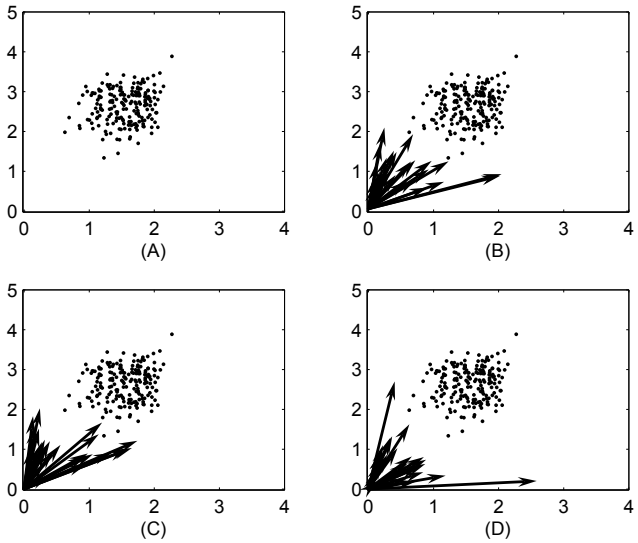

Fig. 6. The business card images plotted in two dimensions to show that data and solutions have pattern like the ones in Figure 1. The x–axis is the an image of the AAU logo, and the y–axis is an image vector capturing the 'hair' region.

shown in Figure 5(B–D). In this simulation the sparse affine NMF algorithm estimates more sparse basis pictures and most basis pictures describe one physical object only.

A two dimensional subspace (axes formed by a picture with 'hair' and an picture with the AAU-logo) of the images in Figure 5 are shown in Figure 6. As above we find that none of the standard NMF's nor sparse NMF basis vectors describe the AAU logo without also capturing 'hair'. The basis pictures for the proposed method however are found close to the axes meaning that they either capture hair or the AAU' logo.

4. DISCUSSION AND CONCLUSION

Non-negative matrix factorization is widely applied because of the ability to create 'parts based' representations, hence, facilitating model interpretation. However, uniqueness is important for the parts based representations to be meaningful. Lack of uniqueness can happen in several ways, e.g., due to an offset vector W_0 as discussed here. Another mechanism resulting in lack of uniqueness is if the support of the process creating a row of H does not include $H = 0$, i.e., if there is an offset in the row variable of H. The H_0 offset can be seen as a W_0 offset with the constraint that W_0 is in the positive span of the column vectors in W

$$R = W(H + H_0\mathbf{1}^T) = WH + W_0\mathbf{1}^T, \;\; W_0 = WH_0 \quad (13)$$

Hence, the H offset issue is a special case of the model we have discussed here: If the resulting W_0 is in the positive span of the columns of W, they can be interpreted as H offsets.

In this work we have defined the augmented non-negative linear mixing model - the sparse affine NMF. We have presented three case stories in which the new sparse affine NMF algorithm outperforms the standard algorithms and a naive solution in estimation of the underlying structure of the data.

5. REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization." *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[2] D. D. Lee and S. H. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2000, pp. 556–562. [Online]. Available: http://citeseer.ist.psu.edu/lee01algorithms.html

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

[4] J. Eggert and E. Körner, "Sparse coding and NMF," in *2004 IEEE International Joint Conference on Neural Networks*, 2004, pp. 2529–2533.

[5] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.