

Determining a Suitable Metric When using Non-negative Matrix Factorization*

David Guillamet and Jordi Vitrià

Computer Vision Center, Dept. Informàtica
Universitat Autònoma de Barcelona
08193 Bellaterra, Barcelona, Spain
{davidg,jordi}@cvc.uab.es

Abstract

The Non-negative Matrix Factorization technique (NMF) has been recently proposed for dimensionality reduction. NMF is capable to produce a region- or part-based representation of objects and images. The positive space defined with NMF lacks of a suitable metric and this paper experimentally compares NMF to Principal Component Analysis (PCA) in the context of classification trying to determine the best distance metric for the NMF. This paper introduces the use of the Earth Mover's Distance (EMD) as a relevant metric that takes into account the positive definition of the NMF bases leading to obtain the best recognition results when the dimensionality of the problem is correctly chosen. PCA and NMF have also been tested under the presence of occlusions and due to its part-based representation, NMF is able to deal with occlusions improving the PCA results.

1. Introduction

Over the past few years, several pattern recognition systems have been proposed based on principal component analysis (PCA) [7, 6, 11, 1, 9, 10]. Although the details vary, these systems can all be described in terms of the same preprocessing and run-time steps. All of them are characterized by the learning of a set of feature vectors and finding a subspace representation that captures the structure of the data. Usually, when calculating the covariance matrix of the problem, eigenvectors are sorted by decreasing eigenvalue only taking the most representative ones which correspond to the directions of maximum variance. Once the subspace is fully-described by a projection matrix, the classification of a new feature vector is accomplished by projecting and finding the nearest training one.

Recently, a new approach for obtaining a linear repre-

sentation of data has been proposed. This new technique, called Non-negative Matrix Factorization (NMF), was used in the work of Lee and Seung [4] to find parts of objects in an unsupervised way. Non-negative Matrix Factorization differs from other methods by its use of non-negativity constraints. Their work was tested with a set of faces [4] and the obtained NMF bases are localized features that correspond with intuitive notions of the parts of faces.

Both methods, PCA and NMF, are simply based on finding a projection matrix used to project new vectors. Since NMF is a recent technique, it does not provide a natural metric to work with its positive projected vectors. Is for this reason that a distance metric must be defined in this positive space in order to work with the projected vectors in an optimal manner.

This paper presents experimental evaluations of traditional distance measures in the context of digit recognition when using PCA and NMF. We have selected the MNIST digit database [3] because it is a well-known database with a large number of training and testing vectors. We have also introduced the Earth Mover's Distance (EMD) [8] distance metric noticing an improvement in the recognition rates when using low dimensional NMF subspaces. We have also considered occlusions in our testing set and NMF has manifested a good behaviour in front of them when the occluded regions are significantly large.

We have to note that the aim of this work is not to obtain the best classifier of the MNIST database. With this current work, we want to show that it is possible to define a metric when using the NMF that can improve the PCA recognition rates using the same training set of digits. And we compare the NMF with PCA because both methods are based on finding a projection space: one of them is claimed to be the best in terms of reconstructed error (PCA) and the other is based on the definition of positive bases.

2. PCA and NMF techniques

Principal Component Analysis (PCA): Due to the high dimensionality of data, similarity and distance metrics are

*This work was supported by IST project IST-1999-20188-CORKINSPECT, sponsored by the European Commission and by Comissionat per a Universitats i Recerca de la Generalitat de Catalunya and Ministerio de Ciencia y Tecnología grant TIC2000-0399-C02-01.

computationally expensive and some compaction of the original data is needed. Principal Component Analysis is an optimal linear dimensionality reduction scheme with respect to the mean squared error (MSE) of the reconstruction. For a set of N training vectors $X = \{x^1, \dots, x^N\}$ the mean ($\mu = \frac{1}{N} \sum_{i=1}^N x^i$) and covariance matrix ($\Sigma = \frac{1}{N} \sum_{i=1}^N (x^i - \mu)(x^i - \mu)^T$) can be calculated. Defining a projection matrix E composed of the K eigenvectors of Σ with highest eigenvalues, the K -dimensional representation of an original, n -dimensional vector x , is given by the projection $y = E^T(x - \mu)$.

Non-Negative Matrix Factorization (NMF): NMF is a method to obtain a representation of data using non-negativity constraints. These constraints lead to a part-based representation because they allow only additive, not subtractive, combinations of the original data [4]. Given an initial database expressed by a $n \times m$ matrix V , where each column is an n -dimensional non-negative vector of the original database (m vectors), it is possible to find two new matrices (W and H) in order to approximate the original matrix $V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^r W_{ia}H_{a\mu}$. The dimensions of the factorized matrices W and H are $n \times r$ and $r \times m$, respectively. Usually, r is chosen so that $(n + m)r < nm$. Each column of matrix W contains a basis vector while each column of H contains the weights needed to approximate the corresponding column in V using the basis from W . In the PCA context, each column of matrix W represents an eigenvector and the factorized matrix of H represent the eigenprojections. In contrast to PCA, NMF does not allow negative entries in the factorized matrices W and H permitting the combination of multiple basis images to represent an object.

In order to estimate the factorization matrices, an objective function has to be defined. A possible objective function is given by $F = \sum_{i=1}^n \sum_{\mu=1}^m [V_{i\mu} \log(WH)_{i\mu} - (WH)_{i\mu}]$. This objective function can be related to the likelihood of generating the images in V from the bases W and encodings H . An iterative approach to reach a local maximum of this objective function is given by the following rules [4]: $W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu}$, $W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}}$, $H_{a\mu} \leftarrow H_{a\mu} \sum_i W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}$. Initialization is performed using positive random initial conditions for matrices W and H . The convergence of the process is also ensured. See [4, 5] for more information.

3. Distance Measures

Four commonly used distance measures are tested in this current work: L1, L2, angle and Earth Mover's Distance (EMD). The angle measure is defined as:

$$\text{dist}(x, y) = -\frac{x \cdot y}{\|x\| \|y\|} \quad (1)$$

and the EMD distance, as known to be the minimal amount of work that must be performed to transform one feature distribution into the other, is based on a solution to the old transportation problem from linear optimization [2]. Let I be a set of suppliers, J a set of consumers and c_{ij} the cost to ship a unit of supply from $i \in I$ to $j \in J$ and usually is defined as the euclidean distance. Now, we want to seek a set of f_{ij} that minimizes the overall cost:

$$\text{dist}(x, y) = \min \sum_{i \in I} \sum_{j \in J} c_{ij} f_{ij} \quad (2)$$

subject to the following constraints:

$$f_{ij} \geq 0 \quad i \in I, j \in J \quad (3)$$

$$\sum_{i \in I} f_{ij} \leq y_j, \quad j \in J \quad (4)$$

$$\sum_{j \in J} f_{ij} \leq x_i, \quad i \in I \quad (5)$$

$$\sum_{i \in I} \sum_{j \in J} f_{ij} = \min \left(\sum_{i \in I} x_i, \sum_{j \in J} y_j \right) \quad (6)$$

Where x_i is the total supply of supplier i and y_j is the total capacity of consumer j . Constraint (3) allows a shipping of supplies from a supplier to a consumer and not vice versa. Constraint (4) forces the consumers to fill up all of their capacities and constraint (5) limits the supply that a supplier can send as a total amount. Constraint (6) forces to move the maximum amount of supplies possible. If the total demands equals the total supply, the distributions have the same overall mass and the EMD is a true metric [8].

4. Experimental Results

We want to test the Non-negative Matrix Factorization (NMF) in a widely used database, the MNIST [3], in order to compare its ability of classification in a well-known pattern recognition problem and compare its recognition rates with the Principal Component Analysis (PCA). There are several methods that have been tested with this digit database and most of them are based on preprocessing the input images in order to reduce some distortion effects. In our case, we have used the 28×28 images of this digit database without any modification.

We have randomly selected 2,000 training vectors (200 of each digit) to learn our PCA and NMF models. The reason of selecting only 2,000 training vectors instead of a large number is that NMF needs to work with all the matrices which are of size $2,000 \times 784 \times 8 = 12\text{Mb}$. We have estimated that having 2,000 training vectors to obtain our NMF model is a good trade-off between time of calculation and accuracy in results.

We have learned both PCA and NMF models and we have obtained a set of bases that can be seen in figures (1)

and (2). Figure (1) shows the bases obtained if we decide to work with a 20 dimensional subspace and figure (2) shows the bases in a 50 dimensional subspace. The main difference between these bases is that NMF obtains a part-based representation, if we work on a high dimensional subspace (50D). In figure (1) this behaviour is not so manifested because we have 10 different digits and we are imposing to obtain 20 bases. With PCA, we can appreciate that if we increase the number of dimensions from 20 to 50, the first 20 are the same and all of them are combinations of global behaviours. But, with NMF, when we increase the number of desired dimensions, we obtain more specific bases corresponding to parts of digits.



Figure 1. Bases obtained by PCA (left) and NMF (right) in a 20D subspace.

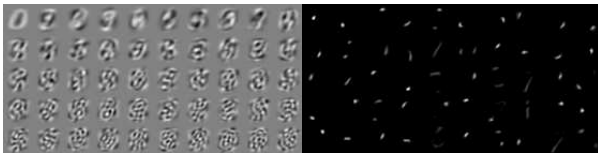


Figure 2. Bases obtained by PCA (left) and NMF (right) in a 50D subspace.

Figure (3) shows two representative NMF calculated bases from figure (2). Because the NMF is based on the maximisation of an objective function, the learned bases can present non-negligible correlations or other higher order effects. In figure (3) we see that both bases are sharing a lot of pixels in common regions. This induces to define a distance between projected vectors that takes into account this fact. EMD is well suited for this problem because we can explicitly define a distance between our bases creating a cost matrix used in the minimization problem defined in expression (2).



Figure 3. 2 different bases obtained with NMF sharing some spatially pixels.

4.1. Recognition without occlusions

In this section we present our experimental results with the testing MNIST images (10,000). We have learned our PCA and NMF models and, in the testing step, we project all the training images (60,000) in our learned model and

given a projected testing image, we search for the most similar training one using a metric in the subspace. Table (1) presents a detailed comparison of all methods. With PCA we have used the well-known L2 metric. NMF has been used with L1, L2, angle and EMD. With EMD, the cost matrix used to weight the transition of one basis (b_i) to another (b_j) is defined as $c_{ij} = dist_{angle}(b_i, b_j)$ where $dist_{angle}$ is the distance defined in expression (1). All techniques have also been tested using a $k - nn$ classifier ($k = 5$).

| Method | 20D | 25D | 30D | 35D | 40D | 45D | 50D | 100D |
|---------------|-------|-------|-------|-------|-------|-------|-------|-------|
| PCA + L2 | 96.60 | 97.21 | 97.42 | 97.56 | 97.34 | 97.43 | 97.49 | 97.21 |
| PCA + L2 5nn | 97.27 | 97.46 | 97.57 | 97.61 | 97.49 | 97.56 | 97.63 | 97.35 |
| NMF + L1 | 93.30 | 94.12 | 95.36 | 95.44 | 95.67 | 95.32 | 95.88 | 96.89 |
| NMF + L1 5nn | 94.15 | 94.86 | 96.14 | 96.27 | 96.36 | 96.31 | 96.56 | 97.10 |
| NMF + L2 | 92.88 | 93.41 | 94.98 | 95.05 | 95.10 | 94.91 | 95.71 | 96.28 |
| NMF + L2 5nn | 93.71 | 93.98 | 95.54 | 95.64 | 95.73 | 95.43 | 96.10 | 96.93 |
| NMF + EMD | 94.32 | 95.43 | 96.74 | 96.43 | 96.56 | 96.12 | 96.08 | 95.78 |
| NMF + EMD 5nn | 95.35 | 96.29 | 97.89 | 97.71 | 97.21 | 96.84 | 96.71 | 96.21 |
| NMF + Cos | 92.84 | 93.27 | 95.06 | 95.32 | 95.11 | 95.49 | 95.72 | 95.91 |
| NMF + Cos 5nn | 89.19 | 89.83 | 91.38 | 91.78 | 91.89 | 91.37 | 91.73 | 93.67 |

Table 1. Results without occlusions.

From our experimental results we can extract several conclusions. The most interesting one is that the combination of NMF+EMD is a good choice when the number of dimensions is not so large. But when the number of dimensions of the subspace is high (100D), L1 is the best and EMD loses its ability of classifying. EMD is based on finding a measure of correlation between bases to define its cost matrix, but if our bases are independent (when we use a high dimensional subspace (100D)), we can not take advantage of this distance. So, if our bases contain some intersecting pixels (until 50D subspace in this problem), EMD is the best metric leading to enhance PCA recognition rates in some particular dimensions (30,35). But when a high dimensional subspace is required, L1 will be the best choice.

4.2. Recognition with occlusions

In this section, two levels of occlusions have been considered. According to the distribution of quadrants that can be seen in figure (4), we have occluded our testing set using one quadrant (25% of a digit area) and two quadrants (50% of a digit area). In figure (4) we can see the reconstruction obtained in three particular examples using two different dimensional spaces (20 and 50). We have occluded quadrants Q2 + Q3 in these examples. As seen in figure (4), PCA always introduces noise in the reconstruction images because it is a global technique even if we are working on a low or high dimensional subspace. NMF also introduces relevant noise but only when we are working on a low dimensional subspace (20) given that the obtained digit bases are not really parts of digits (see figure (1)). In a high dimensional space (50), NMF introduces less noise than PCA because its bases are parts of digits (as seen in figure (2)) and for this reason, NMF is able to classify with the best recognition rates if occlusions are present. Recognition rates of both methods (PCA and NMF) under different levels of occlusions are shown in table (2).

Distribution of quadrants

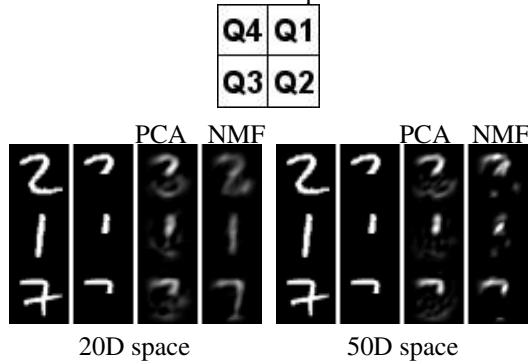


Figure 4. Reconstruction examples with occlusions.

| Results with occlusions in a 20D subspace | | | | | | | | |
|---|-------|-------|-------|-------|-------|-------|-------|-------|
| Method | Q1 | Q2 | Q3 | Q4 | Q1+Q2 | Q2+Q3 | Q3+Q4 | Q4+Q1 |
| PCA + L2 | 87.71 | 85.91 | 84.36 | 86.74 | 53.91 | 60.61 | 55.51 | 57.21 |
| PCA + L2 5nn | 90.87 | 89.25 | 87.69 | 89.52 | 57.54 | 62.19 | 59.44 | 61.56 |
| NMF + L1 | 85.67 | 83.21 | 82.45 | 84.39 | 51.84 | 57.52 | 52.12 | 54.98 |
| NMF + L1 5nn | 88.79 | 86.73 | 84.32 | 87.91 | 55.71 | 59.93 | 56.67 | 58.13 |
| NMF + L2 | 84.56 | 82.34 | 80.12 | 82.10 | 48.65 | 56.71 | 49.26 | 52.31 |
| NMF + L2 5nn | 87.12 | 85.16 | 83.87 | 85.42 | 53.21 | 58.30 | 52.99 | 54.28 |
| NMF + EMD | 90.31 | 88.81 | 85.69 | 86.49 | 55.81 | 61.24 | 56.12 | 57.84 |
| NMF + EMD 5nn | 92.61 | 91.44 | 90.57 | 90.23 | 61.27 | 64.21 | 61.59 | 63.42 |
| NMF + Cos | 91.88 | 90.11 | 87.65 | 87.79 | 57.03 | 62.48 | 59.24 | 59.66 |
| NMF + Cos 5nn | 93.05 | 91.69 | 90.45 | 91.03 | 62.23 | 66.06 | 65.58 | 67.37 |

| Results with occlusions in a 50D subspace | | | | | | | | |
|---|-------|-------|-------|-------|-------|-------|-------|-------|
| Method | Q1 | Q2 | Q3 | Q4 | Q1+Q2 | Q2+Q3 | Q3+Q4 | Q4+Q1 |
| PCA + L2 | 92.21 | 89.48 | 87.76 | 91.26 | 57.82 | 64.95 | 59.16 | 60.39 |
| PCA + L2 5nn | 93.87 | 91.73 | 89.44 | 92.45 | 62.29 | 67.01 | 63.25 | 63.37 |
| NMF + L1 | 86.98 | 83.19 | 82.31 | 85.67 | 54.15 | 55.55 | 52.99 | 47.88 |
| NMF + L1 5nn | 89.45 | 84.41 | 84.09 | 87.02 | 56.21 | 57.64 | 54.54 | 52.71 |
| NMF + L2 | 84.67 | 82.92 | 81.26 | 83.78 | 56.43 | 57.16 | 57.40 | 51.10 |
| NMF + L2 5nn | 86.18 | 85.63 | 83.34 | 86.56 | 59.27 | 58.89 | 60.50 | 56.82 |
| NMF + EMD | 85.39 | 81.54 | 81.07 | 82.36 | 52.75 | 56.12 | 51.32 | 48.13 |
| NMF + EMD 5nn | 88.23 | 83.45 | 83.98 | 84.56 | 55.92 | 57.08 | 53.75 | 49.67 |
| NMF + Cos | 91.49 | 88.68 | 86.67 | 87.22 | 66.48 | 64.59 | 70.52 | 59.89 |
| NMF + Cos 5nn | 94.04 | 90.87 | 91.55 | 90.17 | 69.32 | 68.38 | 72.56 | 65.96 |

Table 2. Results with occlusions in a 20D subspace (above) and in a 50D subspace (below)

From table (2), if a 25% of a digit is occluded, the NMF with the angle distance is a good choice when we are working on low dimensional subspaces (20D). In this case, NMF always improves the PCA recognition rates. However, in a high dimensional space (50D), we find that NMF+angle is the best combination but also, in some particular cases, PCA can give good results. The reason is that PCA is able to reconstruct the original digit because it has enough information about it and, in this case, the additional noise introduced contributes to recover the original digit. In a low dimensional subspace, NMF has the same behaviour as PCA but, in a high dimensional one, it is not able to recover the original digit because its bases are localized parts of digits.

But when a 50% of a digit is occluded, NMF with the angle distance is the best configuration because always outperforms the PCA recognition rates. This is due to the fact that PCA has not enough information about the digits and, usually, when trying to recover them, it fails leading to generate a bad estimation (as seen in the first row of figure (4)). NMF can also add noise to the reconstructed digits but this noise is not so relevant as with the PCA leading to obtain better results. We have to note that NMF+EMD can also provide us some relevant recognition rates if a 50% of a

digit is occluded, but only in low dimensional subspaces.

5. Conclusions

In this paper we have experimentally analysed an alternative technique to Principal Component Analysis (PCA), the so called Non-negative Matrix Factorization (NMF). NMF was initially tested with faces to obtain parts of faces but not used in a classification framework. The results of this paper demonstrate that NMF can be used in a classification framework. NMF is a recent technique and lacks of a suitable metric distance to work with its projected positive vectors. Different distances such as L1, L2 and angle have been tested with the NMF projected vectors. But these distances do not take into account the positive aspects of the NMF. The earth mover's distance (EMD) has been introduced and plays a key role in order to take into account this positive property. When using a low dimensional subspace, the conjunction of NMF and EMD is the best solution improving the classical PCA recognition rates because the obtained NMF bases have intersecting localized pixels. And if a high dimensional subspace is required, the best solution is to use the L1-norm. But the most interesting finding of this paper is that NMF has a good response in front of the presence of occlusions: PCA can not manage with high degrees of occlusions because it is based on a global representation and NMF can improve the PCA recognition rates because it is based on a local representation. For this reason, we believe that NMF can be a relevant technique for pattern recognition problems, where occlusions that can not be handled by PCA may appear.

References

- [1] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. PAMI*, 19(7):720–771, 1997.
- [2] F. Hitchcock. The distribution of a product from several sources to numerous localities. *J. Math. Phys.*, 20:224–230, 1941.
- [3] Y. LeCun. *The MNIST DataBase of Handwritten digits*. <http://yann.lecun.com/exdb/mnist/index.html>.
- [4] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.
- [5] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. *NIPS*, 2000.
- [6] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. PAMI*, 19(7):696–710, 1997.
- [7] H. Murase and S. Nayar. Visual learning and recognition of 3d objects from appearance. *IJCV*, 14:5–24, 1995.
- [8] Y. Rubner, C. Tomasi, and L. Guibas. A metric for distributions with applications to image databases. In *ICCV*, 1998.
- [9] D. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. PAMI*, 18(8):831–836, 1996.
- [10] D. Swets and J. Weng. Hierarchical discriminant analysis for image retrieval. *IEEE Trans. PAMI*, 21(5):386–401, 1999.
- [11] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Neuroscience*, 3(1):71–86, 1991.