Zhijian Yuan and Erkki Oja. A FastICA Algorithm for Non-negative Independent Component Analysis. *In Puntonet, Carlos G.; Prieto, Alberto (Eds.), Proceedings of the Fifth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, Springer Lecture Notes in Computer Science 3195, pp. 1-8, Granada, Spain, 2004.

A FAMILY OF MODIFIED PROJECTIVE NONNEGATIVE MATRIX FACTORIZATION ALGORITHMS

Zhijian Yuan and Erkki Oja

Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 5400, 02015 HUT, Finland
{zhijian.yuan, erkki.oja}@hut.fi

ABSTRACT

We propose here new variants of the Non-negative Matrix Factorization (NMF) method for learning spatially localized, sparse, part-based subspace representations of visual or other patterns. The algorithms are based on positively constrained projections and are related both to NMF and to the conventional SVD or PCA decomposition. A crucial question is how to measure the difference between the original data and its positive linear approximation. Each difference measure gives a different solution.

Several iterative positive projection algorithms are suggested here, one based on minimizing Euclidean distance and the others on minimizing the divergence of the original data matrix and its non-negative approximation. Several versions of divergence such as the Kullback-Leibler, Csiszár, and Amari divergence are considered, as well as the Hellinger and Pearson distances. Experimental results show that versions of P-NMF derive bases which are somewhat better suitable for a localized and sparse representation than NMF, as well as being more orthogonal.

1. INTRODUCTION

For compressing, denoising and feature extraction of high dimensional data such as digital image windows, one of the classical approaches is Principal Component Analysis (PCA) and its extensions and approximations such as the Discrete Cosine Transform. In PCA or the related Singular Value Decomposition (SVD), the data vector is projected on the eigenvectors of the data covariance matrix, each of which provides one linear feature. The representation of a data item in this basis is *distributed* in the sense that typically all the features are used at least to some extent in the reconstruction.

The Non-negative Matrix Factorization (NMF), by Lee and Seung [7], was shown to be a useful technique in approximating high dimensional data where the data are comprised of non-negative components. The authors proposed the idea of using NMF techniques to find a set of basis functions to represent image data where the basis functions enable the identification and classification of intrinsic "parts" that make up the object being imaged by multiple observations. NMF has been typically applied to image and text data, but has also been used to deconstruct music tones.

Recently, the present authors started from the ideas of SVD and NMF and proposed a novel method which we call Projective Non-negative Matrix Factorization (P-NMF), for learning spatially localized, parts-based representations of visual patterns [10]. It can be seen as a combination of ideas from NMF and SVD. It turns out that for P-NMF, as for NMF, there is no unique solution, but the approximation obtained with these techniques depends essentially on the norm or distance measure used. The results obtained using different distance measures vary a lot in the characteristics of the obtained approximation. One such characteristic is sparsity of the representation; this can be objectively measured using entropy. Another characteristic is the degree of orthogonality of the basis vectors obtained. It is the purpose of the present work to take a look at different distance measures and how they effect these two characteristics.

2. THE PROJECTIVE NMF METHOD

Given $m \times n$ nonnegative matrix $\mathbf{V}$, $m < n$, the *Projective Non-negative Matrix Factorization* (P-NMF) is to solve the following optimality problem

$$\min_{\mathbf{W} \geq 0} ||\mathbf{V} - \mathbf{W}\mathbf{W}^T\mathbf{V}||, \tag{1}$$

where $||\cdot||$ is a matrix norm. Note that with the positivity constraint, the orthogonality of $\mathbf{W}$ is not ensured any more, and the method is projective only approximately.

The Projective Non-negative Matrix Factorization (P-NMF) [10] uses only one parameter matrix $\mathbf{W}$ instead of $\mathbf{W}$ and $\mathbf{H}$ in Non-negative Matrix Factorization. The weight matrix $\mathbf{H}$ in NMF is simply replaced by $\mathbf{W}^T\mathbf{V}$ in P-NMF algorithms. The update rules could be obtained similar to Lee and Seung's algorithms [8].

The update rules for Euclidean distance is:

$$\mathbf{W}_{ij} \leftarrow \mathbf{W}_{ij} \frac{2(\mathbf{V}\mathbf{V}^T\mathbf{W})_{ij}}{(\mathbf{W}\mathbf{W}^T\mathbf{V}\mathbf{V}^T\mathbf{W})_{ij} + (\mathbf{V}\mathbf{V}^T\mathbf{W}\mathbf{W}^T\mathbf{W})_{ij}} \tag{2}$$

$$\mathbf{W} \leftarrow \mathbf{W}/norm(\mathbf{W}), \tag{3}$$

1-4244-0779-6/07/$20.00 ©2007 IEEE

and for the divergence measure, the update rule becomes

$$\mathbf{W}_{ij} \leftarrow \mathbf{W}_{ij} \frac{\sum_k \mathbf{V}_{ik}(\mathbf{W}^T\mathbf{V})_{jk}/(\mathbf{W}\mathbf{W}^T\mathbf{V})_{ik}}{\sum_k ((\mathbf{W}^T\mathbf{V})_{jk} + \sum_l \mathbf{W}_{lj}\mathbf{V}_{ik})} + \quad (4)$$

$$\mathbf{W}_{ij} \frac{\sum_k \mathbf{V}_{ik} \sum_l \mathbf{W}_{lj}\mathbf{V}_{lk}/(\mathbf{W}\mathbf{W}^T\mathbf{V})_{lk}}{\sum_k ((\mathbf{W}^T\mathbf{V})_{jk} + \sum_l \mathbf{W}_{lj}\mathbf{V}_{ik})} \quad (5)$$

$$\mathbf{W} \leftarrow \mathbf{W}/norm(\mathbf{W}). \quad (6)$$

P-NMF was shown to work well, especially, it gives a more localized and sparse representation than general NMF algorithms [10]. The question we address here is whether variants of the method can be developed using some other distance or divergence measures, and how would these variants compare with NMF in terms of localized, sparse representations and the orthogonality of the basis vectors.

3. CSISZÁR'S φ-DIVERGENCE

Divergence functions of information measures play an important role in many areas of pattern recognition and machine learning. One of the generalized divergences is given by Csiszár [3], called Csiszár's φ-divergence. In general terms it can be defined as

$$D_C(z||y) = \sum_{k=1}^{N} z_k \varphi(\frac{y_k}{z_k}) \quad (7)$$

where $y_k \geq 0$, $z_k \geq 0$ and $\varphi : [0, \infty) \rightarrow (-\infty, \infty)$ is a function which is convex on $(0, \infty)$ and continuous at zero. To make the Csiszár's φ-divergence a distance measure, we assume that $\varphi(1) = 0$ and it is strictly convex at 1.

Choosing different functions φ gives us many different distance measures, for example:

1. Hellinger distance: If $\varphi(u) = (\sqrt{u} - 1)^2$, then we get the corresponding distance $D_H = \sum_{ik}(\sqrt{y_{ik}} - \sqrt{z_{ik}})^2$.

2. Pearson's distance: If $\varphi(u) = (u-1)^2$, then we get $D_P = \sum_{ik}(y_{ik} - z_{ik})^2/z_{ik}$.

3. Amari's alpha divergences: If $\varphi(u) = u(u^{\beta-1} - 1)/(\beta^2 - \beta) + (1 - u)/\beta$, then we get

$$D_A^{(\beta)}(Z||Y) = \sum_{ik} y_{ik} \frac{(y_{ik}/z_{ik})^{\beta-1} - 1}{\beta(\beta - 1)} + \frac{z_{ik} - y_{ik}}{\beta}. \quad (8)$$

For $\beta \rightarrow 1$ we get the generalized Kullback-Leibler divergence, and for $\beta \rightarrow 0$ the generalized dual Kullback-Leibler divergence.

4. NEW ALGORITHMS

The algorithms for minimizing these distances and preserving positivity follow the same idea as in NMF and P-NMF: starting from a gradient descent, we find a suitable step size such that the algorithms become multiplicative instead of additive. When everything is positive or non-negative initially, this property will be maintained by the multiplicative update rules and a non-negative solution is guaranteed after convergence.

As an example, let us start from Amari's alpha divergence. First, we compute the partial differential of $\mathbf{D}_A^{(\beta)}(\mathbf{W}\mathbf{W}^T\mathbf{V}||\mathbf{V})$ with respect to $\mathbf{W}$

$$\frac{\partial D_A^{(\beta)}(\mathbf{W}\mathbf{W}^T\mathbf{V}||\mathbf{V})}{\partial \mathbf{w}_{ij}} = -\sum_{k,l} \frac{(\mathbf{V}_{kl}/(\mathbf{W}\mathbf{W}^T\mathbf{V})_{kl})^\beta}{\beta} \mathbf{W}_{kj}\mathbf{V}_{il} \quad (9)$$

$$- \sum_{k,l} \frac{(\mathbf{V}_{il}/(\mathbf{W}\mathbf{W}^T\mathbf{V})_{il})^\beta}{\beta} \mathbf{W}_{kj}\mathbf{V}_{kl} \quad (10)$$

$$+ \frac{1}{\beta} \sum_{k,l} (\mathbf{W}_{kj}\mathbf{V}_{il} + \mathbf{W}_{kj}\mathbf{V}_{kl}) \quad (11)$$

Choosing suitable step size,

$$\eta_{ij} = \frac{\mathbf{W}_{ij}}{\sum_{kl} \mathbf{W}_{kj}\mathbf{V}_{il} + \sum_l (\mathbf{W}^T\mathbf{V})_{jl}} \quad (12)$$

we obtain the following algorithm:

$$\mathbf{W}_{ij} \leftarrow \mathbf{W}_{ij} \frac{\sum_{k,l}(\mathbf{V}_{kl}/(\mathbf{W}\mathbf{W}^T\mathbf{V})_{kl})^\beta \mathbf{W}_{kj}\mathbf{V}_{il}}{\sum_{kl} \mathbf{W}_{kj}\mathbf{V}_{il} + \sum_l (\mathbf{W}^T\mathbf{V})_{jl}} \quad (13)$$

$$+ \mathbf{W}_{ij} \frac{\sum_{k,l}(\mathbf{V}_{il}/(\mathbf{W}\mathbf{W}^T\mathbf{V})_{il})^\beta \mathbf{W}_{kj}\mathbf{V}_{kl}}{\sum_{kl} \mathbf{W}_{kj}\mathbf{V}_{il} + \sum_l (\mathbf{W}^T\mathbf{V})_{jl}} \quad (14)$$

Similarly, we can develop multiplicatice algorithms using Hellinger distance, Pearson's distance and the dual Pearson's distance as following:

1. For Hellinger distance

$$\mathbf{D}_H(\mathbf{W}\mathbf{W}^T\mathbf{V}||\mathbf{V}) = \sum_{ik} (\sqrt{(\mathbf{W}\mathbf{W}^T\mathbf{V})_{ik}} - \sqrt{(\mathbf{V})_{ik}})^2, \quad (15)$$

we get the update rule

$$\mathbf{W}_{ij} \leftarrow \mathbf{W}_{ij} \frac{\sum_{k,l} \sqrt{\mathbf{V}_{kl}/(\mathbf{W}\mathbf{W}^T\mathbf{V})_{kl}} \mathbf{W}_{kj}\mathbf{V}_{il}}{\sum_{kl} \mathbf{W}_{kj}\mathbf{V}_{il} + \sum_l (\mathbf{W}^T\mathbf{V})_{jl}} \quad (16)$$

$$+ \mathbf{W}_{ij} \frac{\sum_{k,l} \sqrt{\mathbf{V}_{il}/(\mathbf{W}\mathbf{W}^T\mathbf{V})_{il}} \mathbf{W}_{kj}\mathbf{V}_{kl}}{\sum_{kl} \mathbf{W}_{kj}\mathbf{V}_{il} + \sum_l (\mathbf{W}^T\mathbf{V})_{jl}} \quad (17)$$

2. For Pearson's distance

$$\mathbf{D}_P(\mathbf{W}\mathbf{W}^T\mathbf{V}||\mathbf{V}) = \sum_{ik} \frac{((\mathbf{W}\mathbf{W}^T\mathbf{V})_{ik} - \mathbf{V}_{ik})^2}{(\mathbf{W}\mathbf{W}^T\mathbf{V})_{ik}} \quad (18)$$

The update rule is

$$\mathbf{W}_{ij} \leftarrow \mathbf{W}_{ij} \frac{\sum_{k,l} \frac{(\mathbf{V}_{kl})^2}{((\mathbf{W}\mathbf{W}^T\mathbf{V})_{kl})^2} \mathbf{W}_{kj}\mathbf{V}_{il}}{\sum_{kl} \mathbf{W}_{kj}\mathbf{V}_{il} + \sum_l (\mathbf{W}^T\mathbf{V})_{jl}} \quad (19)$$

$$+ \mathbf{W}_{ij} \frac{\sum_{k,l} \frac{(\mathbf{V}_{il})^2}{((\mathbf{W}\mathbf{W}^T\mathbf{V})_{il})^2} \mathbf{W}_{kj}\mathbf{V}_{kl}}{\sum_{kl} \mathbf{W}_{kj}\mathbf{V}_{il} + \sum_l (\mathbf{W}^T\mathbf{V})_{jl}} \quad (20)$$

3. For Pearson's dual distance

$$\mathbf{D}_{dP}(\mathbf{W}\mathbf{W}^T\mathbf{V}||\mathbf{V}) = \sum_{ik} \frac{((\mathbf{W}\mathbf{W}^T\mathbf{V})_{ik} - \mathbf{V}_{ik})^2}{\mathbf{V}_{ik}} \quad (21)$$

The update rule is

$$\mathbf{W}_{ij} \leftarrow \mathbf{W}_{ij} \times \tag{22}$$

$$\frac{\sum_{kl} \mathbf{W}_{kj}\mathbf{V}_{il} + \sum_{l}(\mathbf{W}^T\mathbf{V})_{jl}}{\sum_{k,l} \frac{2(\mathbf{W}\mathbf{W}^T\mathbf{V})_{kl}}{\mathbf{V}_{kl}}\mathbf{W}_{kj}\mathbf{V}_{il} + \sum_{k,l} \frac{2(\mathbf{W}\mathbf{W}^T\mathbf{V})_{il}}{\mathbf{V}_{il}}\mathbf{W}_{kj}\mathbf{V}_{kl}}. \tag{23}$$

5. P-NMF WITH REGULARIZATION

In the standard NMF, we can impose some additional constraints such as sparsity [4], etc. Kompass generalized a divergence measure for nonnegative matrix factorization with adding regularization terms

$$\mathbf{D}_{Ko}(\mathbf{WH}||\mathbf{V}) = \sum_{ik} \mathbf{V}_{ik} \frac{\mathbf{V}_{ik}^{\beta-1} - (\mathbf{WH})_{ik}^{\beta-1}}{\beta(\beta-1)} \tag{24}$$

$$+ \sum_{ik} (\mathbf{WH})_{ik}^{\beta-1} \frac{(\mathbf{WH})_{ik} - \mathbf{V}_{ik}}{\beta} \tag{25}$$

$$+ \alpha_{\mathbf{W}} f_{\mathbf{W}}(\mathbf{W}) + \alpha_{\mathbf{H}} f_{\mathbf{H}}(\mathbf{H}). \tag{26}$$

where the regularization terms $\alpha_{\mathbf{W}} f_{\mathbf{W}}(\mathbf{W})$ and $\alpha_{\mathbf{H}} f_{\mathbf{H}}(\mathbf{H})$ are used to enforce a certain application depedent characteristic of solutions such as smoothness or sparsity. If we set $\alpha_{\mathbf{W}} = \alpha_{\mathbf{H}} = 0$, it reduce to special Csiszár's φ-divergence case, for example, $\beta = 2$, it simplifies to Euclidean distance; $\beta \to 1$, it tends to Kulback-Leibler divegence.

In P-NMF optimal problem, we simply omit the regularization term on matrix $\mathbf{H}$ in the equation (24):

$$\mathbf{D}_{Ko}(\mathbf{W}\mathbf{W}^T\mathbf{V}||\mathbf{V}) = \sum_{ik} \mathbf{V}_{ik} \frac{\mathbf{V}_{ik}^{\beta-1} - (\mathbf{W}\mathbf{W}^T\mathbf{V})_{ik}^{\beta-1}}{\beta(\beta-1)} \tag{27}$$

$$+ \sum_{ik} (\mathbf{W}\mathbf{W}^T\mathbf{V})_{ik}^{\beta-1} \frac{(\mathbf{W}\mathbf{W}^T\mathbf{V})_{ik} - \mathbf{V}_{ik}}{\beta} + \alpha_{\mathbf{W}} f_{\mathbf{W}}(\mathbf{W}). \tag{28}$$

Use gradient descent and setting the step size to be

$$\frac{\mathbf{W}_{ij}}{\sum_{k,l} \left((\mathbf{W}\mathbf{W}^T\mathbf{V})_{kl}^{\beta-1} \mathbf{W}_{kj}\mathbf{V}_{il} + (\mathbf{W}\mathbf{W}^T\mathbf{V})_{il}^{\beta-1} \mathbf{W}_{kj}\mathbf{V}_{kl} \right)} \tag{29}$$

We have the update rule (30) as shown in the top of the next page.

6. SIMULATION

To see the similarities and differences between the variants introduced above, an experiment was conducted. In this experiment, we employed Lee's and Seung's NMF algorithm [7] and our methods, comparing their performance. We used face images from the MIT-CBCL database as experimental data, and derived the NMF and P-NMF expansions for them. The training data set contains 2429 faces. Each face has $19 \times 19 = 361$ pixels and has been histogram-equalized and normalized so that all pixel values are between 0 and 1.

The basis images for NMF and for the family of P-NMF with dimension 49 are shown in Figure 1. These are the 49 columns of the corresponding matrices $\mathbf{W}$, again shown as 19×19 images. All the basis images for NMF and P-NMF are non-negative.

Figure 2 shows the reconstructions for one of the face images in the NMF, and P-NMF subspaces of dimension $r = 49$. For comparison, also the original face image is shown. Visually, the P-NMF method is comparable to NMF.

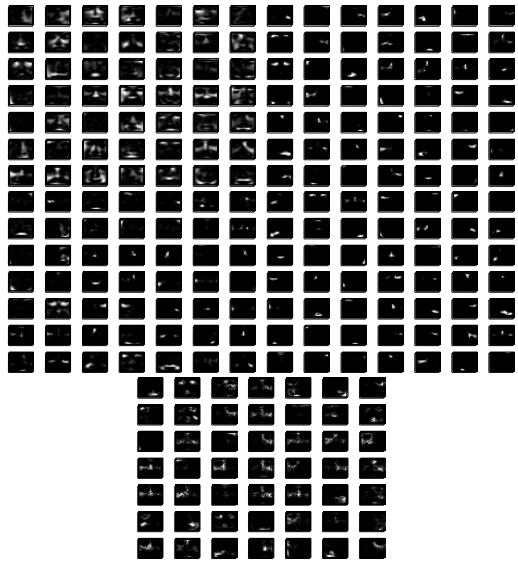

Fig. 1. The bases of face image by Lee and Song's NMF (top, left), P-NMF using KL divergence (top, right), Pearson divergence (middle, left), dual Pearson divergence (middle, right), Hellinger divergence (bottom) with the dimension 49. Each basis component consists of 19×19 pixels.

Fig. 2. The original face image (top, left) and its reconstructions by Lee and Song's NMF (top, middle), P-NMF using KL divergence (top, right), Hellinger divergence (bottom, left), Pearson divergence (bottom, middle), dual Pearson divergence (bottom, right).

To quantify the localization and sparseness, we define entropy as

$$en = -\sum_{1}^{361} p_i \log p_i, \sum_{1}^{361} p_i = 1 \tag{31}$$

where p_i are the elements of the basis images, renormalized so that their sum is equal to 1. It is obvious that smaller entropy value shows more localization and sparseness. Computing the average entropies of the basis matri-

$$\mathbf{W}_{ij} \leftarrow \mathbf{W}_{ij} \frac{\sum_{k,l} \left(\mathbf{V}_{kl} (\mathbf{W}\mathbf{W}^T\mathbf{V})_{kl}^{\beta-2} \mathbf{W}_{kj} \mathbf{V}_{il} + \mathbf{V}_{il} (\mathbf{W}\mathbf{W}^T\mathbf{V})_{il}^{\beta-2} \mathbf{W}_{kj} \mathbf{V}_{kl} \right) - \alpha_{\mathbf{W}} \psi_{\mathbf{W}}(\mathbf{W})}{\sum_{k,l} \left((\mathbf{W}\mathbf{W}^T\mathbf{V})_{kl}^{\beta-1} \mathbf{W}_{kj} \mathbf{V}_{il} + (\mathbf{W}\mathbf{W}^T\mathbf{V})_{il}^{\beta-1} \mathbf{W}_{kj} \mathbf{V}_{kl} \right)} \tag{30}$$

ces derived by NMF, P-NMF with divergence measurement, Hellinger divergence, Pearson divergence and dual Pearson divergence are gives the values 22.329, 7.3534, 6.5044, 7.5338 and 6.2079, respectively. This indicates that the P-NMF bases are markedly sparser than those obtained with NMF. This can also be seen in Fig. 1.

We can also use the orthogonality of the basis vectors as a measure sparseness. The reason is that two nonnegative vectors are orthogonal if and only if they do not share the same non-zero dimensions. Therefore the orthogonality between the learned bases reveals the sparsity of the resulting representations, and the localization for facial images. We measure the orthogonality of the learned bases by the following simple measure

$$\rho = ||W^T W - I||, \tag{32}$$

where $||\cdot||$ refers to the Euclidean matrix norm, and the bases matrix is normalized. Smaller value of ρ indicates higher orthogonality and ρ equals to 0 when the columns of W are completely orthogonal.

Figure (3) compares the orthogonal behavior among a family of P-NMF and NMF. The P-NMF variants converge to local minima with much lower ρ value, that is, they give considerably more orthogonal bases than the original NMF.

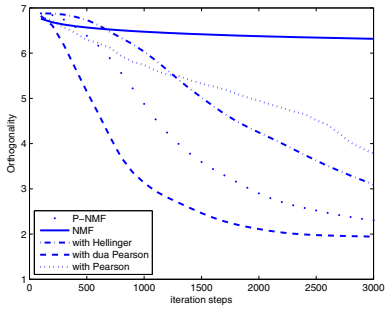

Fig. 3. Orthogonality versus iterative steps using NMF and a family of P-NMF with subdimension 49.

7. CONCLUSION

Projective NMF is a variant of Nonnegative Matrix Factorization (NMF) in which only one parameter matrix is used instead of two matrices. This makes the method somewhat simpler to compute. An open question is what would be the most appropriate distance measure to be used in minimizing the approximation error. Each different distance measure gives a different solution. Here, several relevant distance measures were introduced for the problem, using variants of Csiszár's φ-divergence as the starting point. Multiplicative gradient algorithms were derived for each, which guarantee the positivity of the approximation, when the algorithms are started from positive initial values.

The sparsity of the ensuing solutions was studied and compared experimentally with each other and NMF. As relevant measures of sparsity, the entropy of the non-negative basis vectors as well as their orthogonality were used. It turned out that on both terms, the P-NMF variants produce significantly sparser representations that NMF. Such sparse representations might act as a bridge between statistical and structural pattern recognition.

8. REFERENCES

[1] A. Bell and T. Sejnowski. The "independent components" of images are edge filters. *Vision Research*, 37: 3327–3338, 1997.

[2] A. Cichocki, R. Zdunek, and S. Amari, Csiszar's Divergences for Non-Negative Matrix Factorization: Family of New Algorithms. 6th International Conference on Independent Component Analysis and Blind Signal Separation, Charleston SC, USA, March 5-8, 2006 Springer LNCS 3889, pp. 32-39.

[3] Csiszár I., Information measures: A critical survey. In: Prague Conference on Information Thoery, Academia Prague. Volume A, (1974) 73-86.

[4] P. O. Hoyer Non-negative Matrix Factorization with sparseness constraints Journal of Machine Learning Research 5:1457-1469, 2004.

[5] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, New York, 2001.

[6] R. Kompass. A generalized divergence measure for nonnegative matrix factorization. Neuroinformatics Workshop, Torun, Poland. September 2005.

[7] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[8] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2000.

[9] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network*, 7: 333–339, 1996.

[10] Z. Yuan and E. Oja. Projective nonnegative matrix factorization for image compression and feature extraction. In: Image Analysis (2005) Springer , Berlin, Germany , pp. 333-342.