A Deep Semi-NMF Model for Learning Hidden Representations

George Trigeorgis GEORGE.TRIGEORGIS08@IMPERIAL.AC.UK
Konstantinos Bousmalis K.BOUSMALIS@IMPERIAL.AC.UK
Stefanos Zafeiriou S.ZAFEIRIOU@IMPERIAL.AC.UK
Björn W. Schuller BJOERN.SCHULLER@IMPERIAL.AC.UK
Department of Computing, Imperial College London, United Kingdom

Abstract

Semi-NMF is a matrix factorization technique that learns a low-dimensional representation of a dataset that lends itself to a clustering interpretation. It is possible that the mapping between this new representation and our original features contains rather complex hierarchical information with implicit lower-level hidden attributes, that classical one level clustering methodologies can not interpret. In this work we propose a novel model, Deep Semi-NMF, that is able to learn such hidden representations that allow themselves to an interpretation of clustering according to different, unknown attributes of a given dataset. We show that by doing so, our model is able to learn low-dimensional representations that are better suited for clustering, outperforming Semi-NMF, but also other NMF variants.

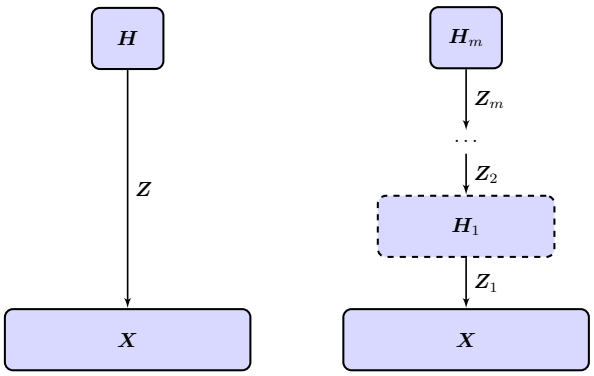
1. Introduction

Matrix factorization is a particularly useful family of techniques in data analysis. In recent years, there has been a significant amount of research on factorization methods that focus on particular characteristics of both the data matrix and the resulting factors. Non-negative matrix factorization (NMF), for example, focuses on the decomposition of non-negative multivariate data matrix $\boldsymbol{X}$ into factors $\boldsymbol{Z}$ and $\boldsymbol{H}$ that are also non-negative, such that $\boldsymbol{X} \approx \boldsymbol{Z}\boldsymbol{H}$. The application area of the family of NMF algorithms has grown significantly during the past years. It has been shown that they can be a successful dimensionality reduction technique over a variety of areas including, but not limited to, environmetrics (Paatero & Tapper, 1994), microarray data analysis (Brunet et al., 2004; Devarajan, 2008), document clustering (Berry & Browne, 2005), face recognition (Zafeiriou et al., 2006; Kotsia et al., 2007) and more. What makes NMF algorithms particularly attractive is the non-negativity constraints imposed on the factors they produce, allowing for better interpretability. Moreover, it has been shown that NMF variants (such as the Semi-NMF) are equivalent to k-means clustering, and that in fact, NMF variants are expected to perform better than k-means clustering particularly when the data is not distributed in a spherical manner (Ding et al., 2010; Cing et al., 2005). Nonlinear extensions of NMF have been also recently studied (Zafeiriou & Petrou, 2010).

In order to extend the applicability of NMF in cases where our data matrix $\boldsymbol{H}$ is not strictly non-negative, Ding et al. (2010) introduced the Semi-NMF, an NMF variant that imposes non-negativity constraints only on the second factor $\boldsymbol{H}$, but allows mixed signs in both the data matrix $\boldsymbol{X}$ and the first factor $\boldsymbol{Z}$. This was motivated from a clustering perspective, where $\boldsymbol{Z}$ represents cluster centroids, and $\boldsymbol{H}$ represents soft membership indicators for every data point, allowing Semi-NMF to learn new lower-dimensional features from the data that have a convenient clustering interpretation.

It is possible that the mapping $\boldsymbol{Z}$ between this new representation $\boldsymbol{H}$ and our original features $\boldsymbol{X}$ contains rather complex hierarchical and structural information. Consider for example the problem of mapping images of faces to their identities: a face image also contains information about attributes like pose and expression that can help identify the person depicted. One could argue that by further factorizing this mapping $\boldsymbol{Z}$, in a way that each factor adds an extra layer of abstraction minimizing the dimensionality of the representation, one could automatically learn such latent attributes and the intermediate hidden representations that are implied, allowing for a better higher-level feature representation $\boldsymbol{H}$, as demonstrated in Figure 1. In this work, we propose a novel Deep Semi-NMF approach, which applies the concept of Semi-NMF to a multi-layer structure that is able to learn hidden representations of the original data. As Semi-NMF has a close relation to k-

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.



(a) Semi-NMF (b) Deep Semi-NMF

Figure 1. (a) A Semi-NMF model results in a linear transformation of the initial input space. (b) Deep Semi-NMF learns a hierarchy of hidden representations that aid in uncovering the final lower-dimensional representation of the data.

means clustering, by extending the model to a deep one we allow our model to learn new representations of our original data that continue to have a clustering interpretation according to the different latent attributes of our dataset, as demonstrated in Figure 2.

Closest to our proposal is recent work that has presented NMF-variants that factorize $\boldsymbol{X}$ into more than 2 factors. Specifically, Ahn et al. (2004) have demonstrated the concept of Multi-Layer NMF on a set of facial images and (Lyu & Wang, 2013; Cichocki & Zdunek, 2006; Song & Lee, 2013) have proposed similar NMF models that can be used for Blind Source Separation, classification of digit images (MNIST), and documents. The representations of the Multi-layer NMF however do not lend themselves to a clustering interpretation, as the representations learned from our model. Although the Multi-layer NMF is a promising technique for learning hierarchies of features from data, we show in this work that our proposed model, the Deep Semi-NMF outperforms the Multi-layer NMF and, in fact, all models we compared it with on the task of feature learning for clustering images of faces.

The novelty of this work can be summarized as follows: *(1)* we outline a novel deep framework for matrix factorization suitable for clustering of multimodally distributed objects such as faces, *(2)* we present a greedy algorithm to optimize the factors of the Semi-NMF problem, inspired by recent advances in deep learning (Hinton & Salakhutdinov, 2006), and *(3)* we evaluate the representations learned by different NMF-variants in terms of clustering performance.

2. Background

In this work, we assume that our data is provided in a matrix form $\boldsymbol{X} \in \mathbb{R}^{p \times n}$, i.e., $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ is a collection of n data vectors as columns, each with p features. Matrix factorization aims at finding factors of $\boldsymbol{X}$ that satisfy certain constraints. In Singular Value Decomposition (SVD) (Golub & Reinsch, 1970), the method that underlies Principal Component Analysis (PCA) (Wold et al., 1987), we factorize $\boldsymbol{X}$ into two factors: the loadings or bases $\boldsymbol{Z} \in \mathbb{R}^{p \times k}$ and the features or components $\boldsymbol{H} \in \mathbb{R}^{k \times n}$, without imposing any sign restrictions on either our data or the resulting factors. In Non-negative Matrix Factorization (NMF) (Lee & Seung, 2001) we assume that all matrices involved contain only non-negative elements[1], so we try to approximate a factorization $\boldsymbol{X}^+ \approx \boldsymbol{Z}^+ \boldsymbol{H}^+$.

2.1. Semi-NMF

In turn, Semi-NMF (Ding et al., 2010) relaxes the non-negativity constrains of NMF and allows the data matrix $\boldsymbol{X}$ and the loadings matrix $\boldsymbol{Z}$ to have mixed signs, while restricting only the features matrix $\boldsymbol{H}$ to comprise of strictly non-negative components, thus approximating the following factorization:

$$\boldsymbol{X}^{\pm} \approx \boldsymbol{Z}^{\pm} \boldsymbol{H}^{+}. \tag{1}$$

This is motivated from a clustering perspective. If we view $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k]$ as the cluster centroids, then $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n]$ can be viewed as the cluster indicators for each datapoint. In fact, if we had a matrix $\boldsymbol{H}$ that was not only non-negative but also orthogonal, then every column vector would have only one positive element, making Semi-NMF equivalent to k-means. Thus Semi-NMF, which does not impose an orthogonality constraint on its features matrix, can be seen as a soft clustering method where the features matrix describes the compatibility of each component with a cluster centroid, a base in $\boldsymbol{Z}$.

2.2. State-of-the-art for learning features for clustering based on NMF-variants

In this work, we compare our method with, among others, the state-of-the-art NMF techniques for learning features for the purpose of clustering. Cai et al. (2011) proposed a graph-regularized NMF (GNMF) which takes into account the intrinsic geometric and discriminating structure of the data space, which is essential to the real-world applications, especially in the area of clustering. To accomplish this, GNMF constructs a nearest neighbor graph to model the manifold structure. By preserving the graph structure,

[1] When not clear from the context we will use the notation $\boldsymbol{A}^+$ to state that a matrix $\boldsymbol{A}$ contains only non-negative elements. Similarly, when not clear, we will use the notation $\boldsymbol{A}^{\pm}$ to state that $\boldsymbol{A}$ may contain any real number.

it allows the learned features to have more discriminating power than the standard NMF algorithm, in cases that the data are sampled from a submanifold which lies in a higher dimensional ambient space. This combines two different notions, that of NMF and graph Laplacian regularization algorithms (Belkin & Niyogi, 2001).

Another state-of-the-art matrix factorization technique would be NeNMF (Guan et al., 2012). NeNMF makes use of Nesterov's optimal gradient method to alternatively optimize one factor, with the other fixed. This allows for faster NMF optimization without time-consuming linesearch procedure or numerical instability problems that traditional NMF has. Guan et al. (2012) showed that it outperformed existing NMF solvers in terms of reconstruction error and document clustering performance.

3. Deep Semi-NMF

In Semi-NMF the goal is to construct a low-dimensional representation $\boldsymbol{H}^+$ of our original data $\boldsymbol{X}^\pm$, with the bases matrix $\boldsymbol{Z}^\pm$ serving as the mapping between our original data and its lower-dimensional representation (see Equation 1). In many cases the data we wish to analyze is often rather complex and has a collection of distinct, often unknown, attributes. In this work for example, we deal with datasets of human faces, where the variability in the data does not only stem from the difference in the appearance of the subjects, but also from other attributes, such as the pose of the head in relation to the camera, or the facial expression of the subject. In addition faces compromise of mainly hierarchical features and thus face clustering problems can be better solved using our deep framework, as each subsequent layers can capture the hierarchical structure.

We propose here the Deep Semi-NMF model, which factorizes a given data matrix $\boldsymbol{X}$ into $m+1$ factors, as follows:

$$\boldsymbol{X}^\pm \approx \boldsymbol{Z}_1^\pm \boldsymbol{Z}_2^\pm \cdots \boldsymbol{Z}_m^\pm \boldsymbol{H}_m^+ \tag{2}$$

This formulation, as shown in with respect to Figures 2 and 1 allows for a hierarchy of m layers of implicit representations of our data that can be given by the following factorizations:

$$\begin{aligned} \boldsymbol{H}_{m-1}^+ &\approx \boldsymbol{Z}_m^\pm \boldsymbol{H}_m^+ \\ &\vdots \\ \boldsymbol{H}_2^+ &\approx \boldsymbol{Z}_3^\pm \cdots \boldsymbol{Z}_m^\pm \boldsymbol{H}_m^+ \\ \boldsymbol{H}_1^+ &\approx \boldsymbol{Z}_2^\pm \cdots \boldsymbol{Z}_m^\pm \boldsymbol{H}_m^+ \end{aligned}$$

As one can see above, we further restrict these implicit representations $(\boldsymbol{H}_1^+, \ldots, \boldsymbol{H}_{m-1}^+)$ to also be non-negative. By doing so, every layer of this hierarchy of representations also lends itself to a clustering interpretation. By examining Figure 2 one can acquire better intuition of how that happens. In this case the input to the model, our data $\boldsymbol{X}$, is a collection of face images from different subjects (identity), expressing a variety of emotions (expressions) taken from many angles (pose). A Semi-NMF model would find a representation $\boldsymbol{H}$ of $\boldsymbol{X}$, which would be useful for performing clustering according to the identity of the subjects, and $\boldsymbol{Z}$ the mapping between these identities and the face images. A Deep Semi-NMF model also finds a representation of our data that has a similar interpretation at the top layer, its last factor $\boldsymbol{H}_m$. However, the mapping from identities to face images is now further analyzed as a product of three factors $\boldsymbol{Z} = \boldsymbol{Z}_1\boldsymbol{Z}_2\boldsymbol{Z}_3$, with $\boldsymbol{Z}_3$ corresponding to the mapping of identities to emotions, $\boldsymbol{Z}_2\boldsymbol{Z}_3$ corresponding to the mapping of identities to poses, and finally $\boldsymbol{Z}_1\boldsymbol{Z}_2\boldsymbol{Z}_3$ corresponding to the mapping of identities to the face images. That means that, as shown in Figure 2 we are able to decompose our data in 3 different ways according to our 3 different attributes:

$$\begin{aligned} \boldsymbol{X}^\pm &\approx \boldsymbol{Z}_1^\pm \boldsymbol{H}_1^+ \\ \boldsymbol{X}^\pm &\approx \boldsymbol{Z}_1^\pm \boldsymbol{Z}_2^\pm \boldsymbol{H}_2^+ \\ \boldsymbol{X}^\pm &\approx \boldsymbol{Z}_1^\pm \boldsymbol{Z}_2^\pm \boldsymbol{Z}_3^\pm \boldsymbol{H}_3^+ \end{aligned}$$

Our hypothesis is that by further factorizing $\boldsymbol{Z}$ we are able to construct a deep model that is able to *(1)* automatically learn what this latent hierarchy of attributes is; *(2)* find representations of the data that are most suitable for clustering according to the attribute that corresponds to each layer in the model; and *(3)* find a better high-level, final-layer representation for clustering according to the attribute with the lowest variability, in our case the identity of the face depicted. In our example in Figure 2 we would expect to find better features for clustering according to identities $\boldsymbol{H}_3$ by learning the hidden representations at each layer most suitable for each of the attributes in our data, in this example: $\boldsymbol{H}_1 \approx \boldsymbol{Z}_2\boldsymbol{Z}_3\boldsymbol{H}_3$ for clustering our original images in terms of poses and $\boldsymbol{H}_2 \approx \boldsymbol{Z}_3\boldsymbol{H}_3$ for clustering the face images in terms of expressions.

In order to expedite the approximation of the factors in our model, we pre-train each of the layers to have an initial approximation of the matrices $\boldsymbol{Z}_i, \boldsymbol{H}_i$ as this greatly improves the training time of the model. This is a tactic that has been employed successfully before (Hinton & Salakhutdinov, 2006) on deep autoencoder networks. To perform the pre-training, we first decompose the initial data matrix $\boldsymbol{X} \approx \boldsymbol{Z}_1\boldsymbol{H}_1$, where $\boldsymbol{Z}_1 \in \mathbb{R}^{p\times k_1}$ and $\boldsymbol{H}_1 \in \mathbb{R}_0^{+\,k_1\times n}$. Following this, we decompose the features matrix $\boldsymbol{H}_1 \approx \boldsymbol{Z}_2\boldsymbol{H}_2$, where $\boldsymbol{Z}_2 \in \mathbb{R}^{k_1\times k_2}$ and $\boldsymbol{H}_1 \in \mathbb{R}_0^{+\,k_2\times n}$, continuing to do so until we have pre-trained all of the layers. Afterwards, we can fine-tune the weights of each layer, by employing alternating minimization (with respect to the objective function in Equation 3) of the two factors in each layer, in order to reduce the total reconstruction error of the model, according to the cost

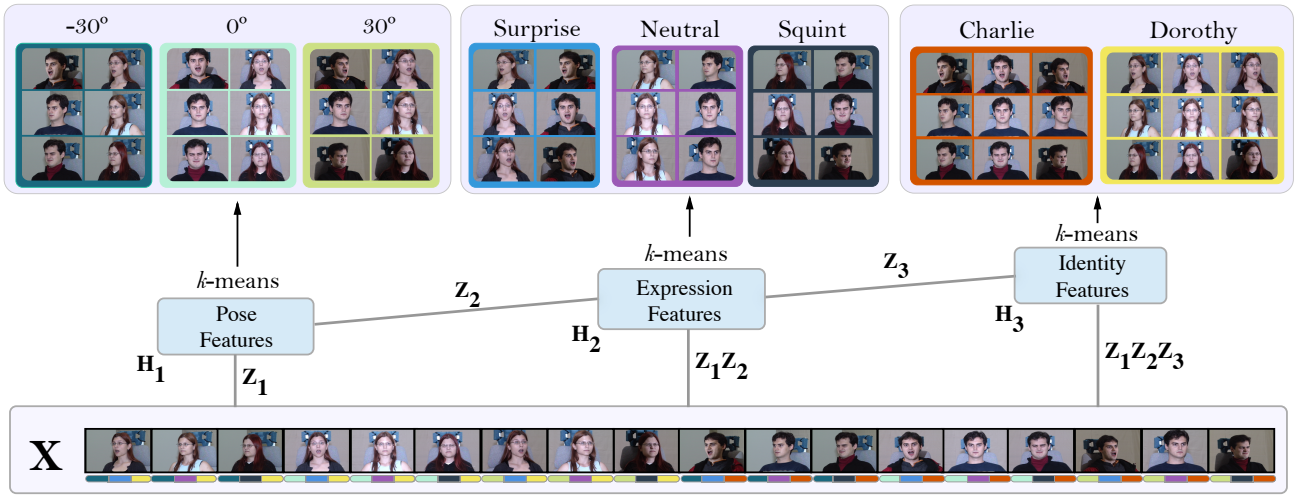


Figure 2. A Deep Semi-NMF model learns a hierarchical structure of features, with each layer learning a representation suitable for clustering according to the different attributes of our data. In this simplified, for demonstration purposes, example from the CMU Multi-PIE database, a Deep Semi-NMF model is able to simultaneously learn features for pose clustering ($\boldsymbol{H}_1$), for expression clustering ($\boldsymbol{H}_2$), and for identity clustering ($\boldsymbol{H}_3$). Each of the images in $\mathbf{X}$ has an associated color coding that indicates its memberships according to each of these attributes (pose/expression/identity).

function in Equation 3.

$$\begin{aligned} C_{\text{deep}} &= \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{Z}_1\boldsymbol{Z}_2\cdots\boldsymbol{Z}_m\boldsymbol{H}_m\|_F^2 \\ &= tr[\boldsymbol{X}^\top\boldsymbol{X} - 2\boldsymbol{X}^\top\boldsymbol{Z}_1\boldsymbol{Z}_2\cdots\boldsymbol{Z}_m\boldsymbol{H}_m \\ &\quad + \boldsymbol{H}_m^\top\boldsymbol{Z}_m^\top\boldsymbol{Z}_{m-1}^\top\cdots\boldsymbol{Z}_1^\top\boldsymbol{Z}_1\boldsymbol{Z}_2\cdots\boldsymbol{Z}_m\boldsymbol{H}_m] \end{aligned} \quad (3)$$

Update rule for the weights matrix $\boldsymbol{Z}$ We fix the rest of the weights for the i_{th} layer and we minimize the cost function with respect to $\boldsymbol{Z}_i$. That is, we set $\partial C_{\text{deep}}/\partial\boldsymbol{Z}_i = 0$, which gives us the updates:

$$\begin{aligned} \boldsymbol{Z}_i &= (\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top\boldsymbol{X}\tilde{\boldsymbol{H}}_i^\top(\tilde{\boldsymbol{H}}_i\tilde{\boldsymbol{H}}_i^\top)^{-1} \\ \boldsymbol{Z}_i &= \boldsymbol{\Psi}^\dagger\boldsymbol{X}\tilde{\boldsymbol{H}}_i^\dagger \end{aligned} \quad (4)$$

where $\boldsymbol{\Psi} = \boldsymbol{Z}_1\cdots\boldsymbol{Z}_{i-1}$, $\dagger$ denotes the Moore-Penrose pseudo-inverse and $\tilde{\boldsymbol{H}}_i$ is the reconstruction of the i^{th} layer's feature matrix.

Update rule for features matrix $\boldsymbol{H}$ Utilizing a similar proof to Ding et al. (2010), we can formulate the update rule for $\boldsymbol{H}_i$ as follows:

$$\boldsymbol{H}_i = \boldsymbol{H}_i \odot \sqrt{\frac{[\boldsymbol{\Psi}^\top\boldsymbol{X}]^{\text{pos}} + [\boldsymbol{\Psi}^\top\boldsymbol{\Psi}]^{\text{neg}}\boldsymbol{H}_i}{[\boldsymbol{\Psi}^\top\boldsymbol{X}]^{\text{neg}} + [\boldsymbol{\Psi}^\top\boldsymbol{\Psi}]^{\text{pos}}\boldsymbol{H}_i}} \quad (5)$$

Supplementary material including the implementation of Algorithm 1 and the proof of its convergence can be found at http://trigeorgis.com/deepseminmf.

Complexity

The computational complexity for the pre-training stage of Deep Semi-NMF is of order $\mathcal{O}\left(mt\left(pnk + nk^2 + kp^2 + kn^2\right)\right)$, where m is the number of layers, t the number of iterations until convergence and k is the maximum number of components out of all the layers. The complexity for the fine-tuning stage is $\mathcal{O}\left(mt_f\left(pnk + (p+n)k^2\right)\right)$ where t_f is the number of additional iterations needed.

Non-linear Update Rules

One can use a non-linear function $g(\cdot)$, between each of the implicit representations $(\boldsymbol{H}_1^+, \ldots, \boldsymbol{H}_{m-1}^+)$, in order to better approximate the non-linear manifolds which the given data matrix $\boldsymbol{X}$ originally lies on. Thus, one can represent the i^{th} feature matrix $\boldsymbol{H}_i$, by

$$g(\boldsymbol{H}_i) \approx \boldsymbol{Z}_{i+1}\boldsymbol{H}_{i+1}. \quad (6)$$

This, creates new derivatives update rules as the derivatives with respect to the objective function become:

$$\frac{\partial\boldsymbol{C}_{deep}}{\partial\boldsymbol{Z}_i} = (\boldsymbol{N}_i - \boldsymbol{R}_i)\boldsymbol{H}_i^\top \quad (7)$$

$$\frac{\partial\boldsymbol{C}_{deep}}{\partial\boldsymbol{H}_i} = \boldsymbol{Z}_i^\top(\boldsymbol{N}_i - \boldsymbol{R}_i) \quad (8)$$

Algorithm 1 Suggested algorithm for training a Deep Semi-NMF model. Initially we approximate the factors greedily using the SEMI-NMF algorithm (Ding et al., 2010) and we fine-tune the factors until we reach the convergence criterion.

function DEEPSEMINMF
Input: $\boldsymbol{X} \in \mathbb{R}^{p \times n}$, list of layer sizes
Output: weight matrices $\boldsymbol{Z}_i$ and feature matrices $\boldsymbol{H}_i$ for each of the layers

Initialize Layers
for all layers **do**
$\boldsymbol{Z}_i, \boldsymbol{H}_i \leftarrow \text{SEMINMF}(\boldsymbol{H}_{i-1}, \text{layers}(i))$
end for

repeat
for all layers **do**
$$\tilde{\boldsymbol{H}}_i \leftarrow \begin{cases} \boldsymbol{H}_i & \textbf{if} \quad i = k \\ \boldsymbol{Z}_{i+1}\tilde{\boldsymbol{H}}_{i+1} & \text{otherwise} \end{cases}$$
$\boldsymbol{\Psi} \leftarrow \prod_{k=1}^{i-1} \boldsymbol{Z}_k$
$\boldsymbol{Z}_i \leftarrow \boldsymbol{\Psi}^{\dagger}\boldsymbol{X}\tilde{\boldsymbol{H}}_i^{\dagger}$
$$\boldsymbol{H}_i \leftarrow \boldsymbol{H}_i \odot \sqrt{\frac{[\boldsymbol{\Psi}^{\top}\boldsymbol{X}]^{\text{pos}} + [\boldsymbol{\Psi}^{\top}\boldsymbol{\Psi}]^{\text{pos}}\boldsymbol{H}_i}{[\boldsymbol{\Psi}^{\top}\boldsymbol{X}]^{\text{neg}} + [\boldsymbol{\Psi}^{\top}\boldsymbol{\Psi}]^{\text{pos}}\boldsymbol{H}_i + \epsilon}}$$
end for
until Stopping criterion is reached
end function

which gives the following update rules:

$$\boldsymbol{Z}_i = \begin{cases} \boldsymbol{X}\tilde{\boldsymbol{H}}_i^{\dagger}; & \textbf{if} \quad i = 1 \\ g(\boldsymbol{H}_{i-1})\tilde{\boldsymbol{H}}_i^{\dagger} & \text{otherwise} \end{cases}, \tag{9}$$

$$\boldsymbol{H}_i = \boldsymbol{H}_i \odot \frac{[\boldsymbol{Z}_i^{\top}\boldsymbol{R}_i]^{\text{pos}} + [\boldsymbol{Z}_i^{\top}\boldsymbol{N}_i]^{\text{neg}}}{[\boldsymbol{Z}_i^{\top}\boldsymbol{R}_i]^{\text{neg}} + [\boldsymbol{Z}_i^{\top}\boldsymbol{N}_i]^{\text{pos}}} \tag{10}$$

where, $\forall i . i \leq m$ $\boldsymbol{N}_i, \boldsymbol{R}_i$ are auxiliary matrices that facilitate the computation of the factors in each layer, $\boldsymbol{A}^{\text{pos}}$ is a matrix that has the negative elements of matrix $\boldsymbol{A}$ replaced with 0, and similarly $\boldsymbol{A}^{\text{neg}}$ is one that has the positive elements of $\boldsymbol{A}$ replaced with 0. The base case values $i = 1$ for these matrices are:

$$\boldsymbol{N}_1 = (\boldsymbol{Z}_1\tilde{\boldsymbol{H}}_1) \odot \nabla g^{-1}(\boldsymbol{Z}_1\boldsymbol{H}_1), \tag{11}$$

$$\boldsymbol{R}_1 = \boldsymbol{X} \odot \nabla g^{-1}(\boldsymbol{Z}_1\boldsymbol{H}_1), \tag{12}$$

and for the cases where $i > 1$ they are computed as such:

$$\boldsymbol{N}_{i+1} = (\boldsymbol{Z}_i^{\top}\boldsymbol{N}_i) \odot \nabla g^{-1}(\boldsymbol{Z}_i\boldsymbol{H}_i), \tag{13}$$

$$\boldsymbol{R}_{i+1} = (\boldsymbol{Z}_i^{\top}\boldsymbol{R}_i) \odot \nabla g^{-1}(\boldsymbol{Z}_i\boldsymbol{H}_i). \tag{14}$$

4. Experiments

Our main hypothesis is that a Deep Semi-NMF is able to learn better high-level representations of our original data than an one-layer Semi-NMF for clustering according to the attribute with the lowest variability in the dataset. In order to evaluate this hypothesis, we have compared the performance of Deep Semi-NMF with that of other methods, on the task of clustering images of faces in 3 distinct datasets. These datasets are:

- **CMU MultiPIE:** The first dataset we examine is *The CMU Multi Pose, Illumination, and Expression* (MultiPIE) Database (Gross et al., 2010) and contains around $750,000$ images of 337 subjects, captured under laboratory conditions in four different sessions. In this work, we used a subset of $13,230$ images of 147 subjects in 5 different poses and 3 different illumination conditions, expressing 6 different emotions. Using the annotations from (Sagonas et al., 2013a;b), we aligned these images based on a common frame by using piece-wise affine warping. After that, we resized them to a smaller resolution of 30×30. The database comes with labels for each of the attributes mentioned above: identity, illumination, pose, expression.
- **CMU PIE:** We also used a freely available version of CMU Pie (Sim et al., 2003), which comprises of $2,856$ grayscale 32×32 face images of 68 subjects. Each person has 42 facial images under different light and illumination conditions. In this database we only know the identity of the face in each image and we could not use piece-wise affine warping, since we did not have facial point annotations for it.
- **XM2VTS:** *The Extended Multi Modal Verification for Teleservices and Security applications* (XM2VTS) (Messer et al., 1999) contains $2,360$ frontal images of 295 different subjects. Each subject has two available images for each of the four different laboratory sessions, for a total of 8 images. The images were aligned based on the same piece-wise affine warping technique as the Multi-PIE dataset, after resizing the original images to 42×30.

In order to evaluate the performance of our model, we compared it against not only Semi-NMF (Ding et al., 2010), but also against other NMF variants that could be useful in learning such representations. More specifically, for each of our three datasets we performed the following experiments:

- **Pixel Intensities:** By using only the pixel intensities of the images in each of our datasets, which of course give us a strictly non-negative input data matrix $\boldsymbol{X}$, we compare the reconstruction error and the clustering performance of our Deep Semi-NMF method against the Semi-NMF, NMF with multiplicative update rules (Lee & Seung, 2001), Multi-Layer NMF (Song & Lee,

Name	Number of components				
	30	40	50	60	70
CMU Multi-PIE – Pixel Intensities					
NMF	2.68	2.52	2.41	2.32	2.23
NeNMF	2.40	2.20	2.04	1.92	1.80
GNMF	3.05	3.03	2.98	3.02	3.02
Semi-NMF	2.37	2.16	2.01	1.89	1.78
Multi-layer NMF	2.71	2.54	2.42	2.31	2.22
Deep Semi-NMF	2.40	2.20	2.07	1.96	1.86
CMU Multi-PIE – Image Gradient Orientations					
Semi-NMF	0.14	0.13	0.13	0.12	0.12
Deep Semi-NMF	0.14	0.13	0.13	0.13	0.12

Table 1. The mean reconstruction error for each of the algorithms with a variable number of components. The error quantifies the deviation of the reconstruction with the original matrix $\boldsymbol{X}$, using the Euclidean objective function we used: $\frac{1}{N}\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\|^2$, with N amount of samples. Deep Semi-NMF has a comparable reconstruction error to that of Semi-NMF.

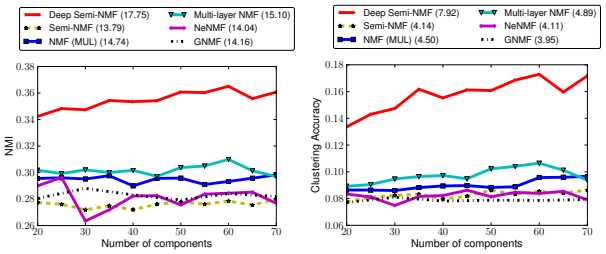
2013), GNMF (Cai et al., 2011), and NeNMF (Guan et al., 2012).

- **Image Gradient Orientations (IGO):** In general, the trend in Computer Vision is to use complicated engineered features like HoGs, SIFT, LBPs, etc. As a proof of concept, we choose to conduct experiments with simple gradient orientations (Zafeiriou et al., 2012) as features, instead of pixel intensities, which results into a data matrix of mixed signs, and expect that we can learn better data representations for clustering faces according to identities. In this case, we only compared our Deep Semi-NMF with its one-layer Semi-NMF equivalent, as the other techniques are not able to deal with mixed-sign matrices.

Finally, we evaluated our secondary hypotheses, i.e. that every hidden representation in each layer is in fact most suited for clustering according to the attributes that corresponds to the layer of interest. We performed clustering experiments by using the features learned in each layer of a two-layer Deep Semi-NMF on the case of CMU Multi-PIE, as this was the only dataset for which we had labels for attributes other than the identity.

4.1. Implementation Details

In order to initiate the matrix factorization process, NMF and Semi-NMF algorithms start from some initial point $(\boldsymbol{Z}^0, \boldsymbol{H}^0)$, where usually $\boldsymbol{Z}^0$ and $\boldsymbol{H}^0$ are randomly initialized matrices. In order to speed up the convergence rate of NMF, Boutsidis & Gallopoulos (2008), suggested Non-negative Double Singular Value decomposition (NNDSVD), which is a new method based on two SVD



(a) NMI (b) Accuracy

Figure 3. **CMU MultiPIE–Pixel Intensities:** NMI and Accuracy for clustering based on the representations learned by each model with respect to identities. The deep architectures are compromised of 2 representation layers 1988-625-a and the representations used were from the top layer. In parenthesis we show the AUC scores.

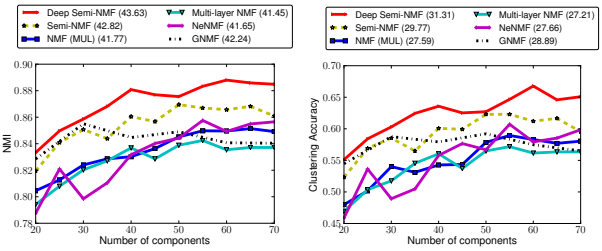
processes, one to approximate the initial data matrix $\boldsymbol{X}$ and the other to approximate the positive sections of the resulting partial SVD factors. Although, the proposed initialization of Semi-NMF by its authors is by using the k-means algorithm (Ding et al., 2010), we found empirically that it was too computationally heavy when the number of components k was fairly high ($k > 100$). As an alternative we used NNDSVD to gain an initial and deterministic approximation, after forcing the initial data to have non-negative values, by setting all negative values to zero.

For the GNMF experimental setup, we chose a suitable number of neighbours, in our case 5, using visualization of the datasets using Laplacian Eigenmaps (Belkin & Niyogi, 2001), such that we had visually distinct clusters.

Important for the experimental setup is the selected structure of the multi-layered models. After careful preliminary experimentation, we focused on experiments that involve two hidden layer architectures for the Deep Semi-NMF and Multi-layer NMF. Although we experimented with a higher-number of layers, approximating the additional number of factors/layers did not seem to have a significant impact on our results for these datasets, and was significantly more computationally expensive. We specifically experimented with models that had a first hidden representation $\boldsymbol{H}_1$ with 625 features, and a second representation $\boldsymbol{H}_2$ with a number of features that ranged from 20 to 70. Again we chose these numbers after preliminary results showed us that the computational burden of additional features outweighed the performance increase obtained by increasing these numbers.

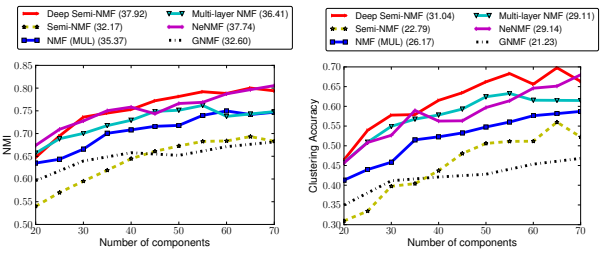
4.2. Reconstruction Error Results

Our first experiment was to evaluate whether the extra layers, which naturally introduce more factors and are therefore more difficult to optimize, result in a lower quality lo-



(a) NMI (b) Accuracy

Figure 4. **XM2VTS-Pixel Intensities:** NMI and Accuracy for clustering based on the representations learned by each model with respect to identities. The deep architectures are compromised of 2 representation layers 1260-625-a and the representations used were from the top layer. In parenthesis we show the AUC scores.



(a) NMI (b) Accuracy

Figure 5. **CMU PIE–Pixel Intensities:** NMI and Accuracy for clustering based on the representations learned by each model with respect to identities. The deep architectures are compromised of 2 representation layers 1024-625-a and the representations used were from the top layer. In parenthesis we show the AUC scores.

cal optimum. We evaluated how well the matrix decomposition is performed by calculating the reconstruction error, the Frobenius norm of the difference between the original data and the reconstruction for all the methods we compared. Note that, in order to have comparable results, all of the methods have the same stopping criterion rules. We have set the maximum amount of iterations to 300 (usually $\sim$100 epochs are enough) and we use the convergence rule $E_{i-1} - E_i \leq \kappa \max(1, E_{i-1})$ in order to stop the process when reconstruction error (E_i) between the current and previous update is small enough. In our experiments we set $\kappa = 10^{-6}$. Table 1 shows the change in reconstruction error with respect to the selected number of features in $\boldsymbol{H}_2$ for all the methods we used on the Multi-PIE dataset. The results for the other datasets and for a larger variety of number of components were similar and are excluded due to lack of space.

The results show that Semi-NMF and Deep Semi-NMF manage to reach a much lower reconstruction error than

	Pose Accuracy (%)		Identity Accuracy (%)	
#	$\boldsymbol{H}_1$	$\boldsymbol{H}_2$	$\boldsymbol{H}_1$	$\boldsymbol{H}_2$
1	27.58	23.11	9.37	16.59
2	28.18	23.25	9.48	16.67

Table 2. Clustering accuracy according to pose and identity labels on the CMU Multi-PIE dataset using two Deep Semi-NMF models with two hidden layers each *(1)* 625-60 and *(2)* 625-70.

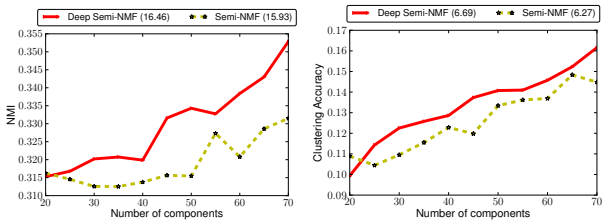
the rest consistently, which would match our expectations as they do not constrain the weights $\boldsymbol{Z}$ to be non-negative. What is important to note here is that the Deep Semi-NMF models do not have a significantly lower reconstruction error compared to the equivalent Semi-NMF models, even though the approximation involves more factors. This is in contrast to the multi-layer NMF and GNMF which sacrifice the reconstruction quality, in return for uncovering more meaningful features than their NMF counterpart.

4.3. Clustering Results

After achieving satisfactory reconstruction error for our method, we proceeded to evaluate the features learned at the final representation layer, by using k-means clustering, following the experimental protocol in (Cai et al., 2011). To assess the clustering quality of the representations produced by each of the algorithms we compared, we take advantage of the fact that the datasets are already labelled. The two metrics used were the accuracy (AC) and the normalized mutual information metric (NMI), as those are defined in (Xu et al., 2003).

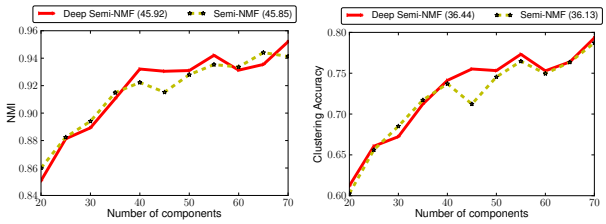
Figures 3-5 show the comparison in clustering accuracy and NMI when using k-means on the feature representations produced by each of the techniques we compared, when our input matrix contained only the pixel intensities of each image. Our method significantly outperforms every method we compared it with on all three datasets. The difference is not as obvious in Figure 5, which could perhaps be a sign that our method is able to use the warping technique we used to its advantage more so than the other techniques we used.

By using IGOs, the Deep Semi-NMF was able to outperform the single-layer Semi-NMF as shown in Figures 6-8. Making use of these simple mixed-signed features improved the clustering accuracy considerably, especially for XM2VTS and CMU PIE. It should be noted that in all cases, with the exception of the XM2VTS experiment with IGOs, our Deep Semi-NMF outperformed all other methods with a difference in performance that is statistically significant (paired t-test, $p \ll 0.01$).



(a) NMI (b) Accuracy

Figure 6. **CMU MultiPIE-IGO:** NMI and Accuracy scores on clustering based on the representations learned by each model with respect to identities. The deep architectures are compromised of 2 hidden layers 1988-625-a and the representations used were from the top layer. In parenthesis we show the AUC scores.

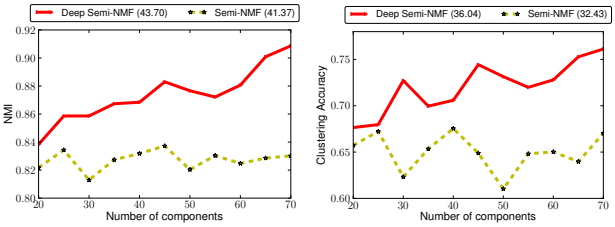


(a) NMI (b) Accuracy

Figure 7. **XM2VTS-IGO:** NMI and Accuracy scores on clustering based on the representations learned by each model with respect to identities. The deep architectures are compromised of 2 hidden layers 2520-625-a and the representations used were from the top layer. In parenthesis we show the AUC scores.

4.4. Clustering with Respect to Different Attributes

Finally, we conducted experiments by performing k-means clustering on each of the two representations learned by our Deep Semi-NMF models by using the raw intensities of our warped images of CMU Multi-PIE. We only used Multi-PIE since we only had identity labels for our other datasets. Since the warping technique we use gets rid of most of the variability shown in expressions, we evaluated how well we did on clustering according to pose and identities. As one can see in Table 2 our first layer indeed learns representations that are better suited for clustering according to poses. On the other hand, we do confirm that our final layer indeed learns representations that are better suited for clustering according to identities, thus confirming our secondary hypothesis, i.e. that every hidden representation in each layer is in fact most suited for clustering according to the attributes that corresponds to the layer of interest. As one can see in Figure 3, by learning the hidden representation that relates to poses, we are able to achieve significantly better results compared to the Semi-NMF, where we learn only one level of representation.



(a) NMI (b) Accuracy

Figure 8. **CMU PIE-IGO:** NMI and Accuracy scores on clustering based on the representations learned by each model with respect to identities. The deep architectures are compromised of 2 hidden layers 2048-625-a and the representations used were from the top layer. In parenthesis we show the AUC scores.

5. Conclusion

We have introduced a novel deep architecture for semi-non-negative matrix factorization, the Deep Semi-NMF , that is able to automatically learn a hierarchy of attributes of a given dataset, as well as representations suited for clustering according to these attributes. We have also presented an algorithm for optimizing the factors of our Deep Semi-NMF, and we evaluate its performance compared to the single-layered Semi-NMF and other related work, on the problem of clustering faces with respect to their identities. We have shown that our technique is able to learn a high-level, final-layer representation for clustering with respect to the attribute with the lowest variability in the case of three popular datasets of face images, outperforming the other NMF-based techniques.

The next obvious step is to explore the suitability of such learned hidden representations for performing classification. Another line of work that will aid the learning process will be the initialization scheme for pre-training the Deep Semi-NMF model, as currently we used a rough approximation of the initial $\boldsymbol{Z}^0, \boldsymbol{H}^0$ matrices using the NNDSVD algorithm. Finally, future avenues include experimenting with other applications, e.g. in the area of speech recognition, especially for multi-source speech recognition and we will investigate multilinear extensions of the proposed framework (Zafeiriou, 2009b;a).

Acknowledgements

We would like to thank our anonymous reviewers for their useful comments. George Trigeorgis is a recipient of the fellowship of the Department of Computing, Imperial College London, and this work was partially funded by it. The work of Konstantinos Bousmalis was funded partially from the Google Europe Fellowship in Social Signal Processing. The work of Stefanos Zafeiriou was partially funded by the EPSRC project EP/J017787/1 (4D-FAB).

References

Ahn, Jong-Hoon, Choi, Seungjin, and Oh, Jong-Hoon. A multiplicative up-propagation algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 3. ACM, 2004.

Belkin, Mikhail and Niyogi, Partha. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pp. 585–591, 2001.

Berry, Michael W and Browne, Murray. Email surveillance using non-negative matrix factorization. *Computational & Mathematical Organization Theory*, 11(3):249–264, 2005.

Boutsidis, Christos and Gallopoulos, Efstratios. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.

Brunet, Jean-Philippe, Tamayo, Pablo, Golub, Todd R, and Mesirov, Jill P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.

Cai, Deng, He, Xiaofei, Han, Jiawei, and Huang, Thomas S. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560, 2011.

Cichocki, Andrzej and Zdunek, Rafal. Multilayer nonnegative matrix factorization. *Electronics Letters*, 42:947–948, 2006.

Cing, C., He, X., and Simon, H.D. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proc. SIAM Data Mining*, 2005.

Devarajan, Karthik. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS computational biology*, 4(7):e1000029, 2008.

Ding, Chris HQ, Li, Tao, and Jordan, Michael I. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55, 2010.

Golub, Gene H and Reinsch, Christian. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14 (5):403–420, 1970.

Gross, Ralph, Matthews, Iain, Cohn, Jeffrey, Kanade, Takeo, and Baker, Simon. Multi-pie. *Image and Vision Computing*, 28(5): 807–813, 2010.

Guan, Naiyang, Tao, Dacheng, Luo, Zhigang, and Yuan, Bo. Nenmf: an optimal gradient method for nonnegative matrix factorization. *Signal Processing, IEEE Transactions on*, 60(6): 2882–2898, 2012.

Hinton, Geoffrey E and Salakhutdinov, Ruslan R. Reducing the dimensionality of data with neural networks. *Science*, 313 (5786):504–507, 2006.

Kotsia, Irene, Zafeiriou, Stefanos, and Pitas, Ioannis. A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems. *IEEE Transactions on Information Forensics and Security*, 2(3-2): 588–595, 2007.

Lee, Daniel D. and Seung, H. Sebastian. Algorithms for nonnegative matrix factorization. *Advances in neural information processing systems*, 13:556–562, 2001.

Lyu, Siwei and Wang, Xin. On algorithms for sparse multi-factor nmf. In *Advances in Neural Information Processing Systems*, pp. 602–610, 2013.

Messer, Kieron, Matas, Jiri, Kittler, Josef, Luettin, Juergen, and Maitre, Gilbert. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pp. 965–966. Citeseer, 1999.

Paatero, Pentti and Tapper, Unto. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pp. 397–403, Dec 2013a.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. A semi-automatic methodology for facial landmark annotation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pp. 896–903, June 2013b.

Sim, Terence, Baker, Simon, and Bsat, Maan. "the cmu pose, illumination, and expression database.". *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003.

Song, Hyun Ah and Lee, Soo-Young. Hierarchical data representation model - multi-layer nmf. *International Conference on Learning Representations*, abs/1301.6316, 2013.

Wold, Svante, Esbensen, Kim, and Geladi, Paul. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1):37–52, 1987.

Xu, Wei, Liu, Xin, and Gong, Yihong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 267–273. ACM, 2003.

Zafeiriou, Stefanos. Algorithms for nonnegative tensor factorization. In *Tensors in Image Processing and Computer Vision*, pp. 105–124. Springer, 2009a.

Zafeiriou, Stefanos. Discriminant nonnegative tensor factorization algorithms. *Neural Networks, IEEE Transactions on*, 20 (2):217–235, 2009b.

Zafeiriou, Stefanos and Petrou, Maria. Nonlinear non-negative component analysis algorithms. *Image Processing, IEEE Transactions on*, 19(4):1050–1066, 2010.

Zafeiriou, Stefanos, Tefas, Anastasios, Buciu, Ioan, and Pitas, Ioannis. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. *Neural Networks, IEEE Transactions on*, 17(3):683–695, 2006.

Zafeiriou, Stefanos, Tzimiropoulos, Georgios, Petrou, Maria, and Stathaki, Tania. Regularized kernel discriminant analysis with a robust kernel for face recognition and verification. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(3): 526–534, 2012.