Anatomical Parts-Based Regression Using Non-Negative Matrix Factorization

Swapna Joshi, S. Karthikeyan, B.S. Manjunath
Department of Electrical and Computer Engineering,
University of California Santa Barbara
{sjoshi,karthikeyan,manj}@ece.ucsb.edu
Scott Grafton
Department of Psychology, University of California Santa Barbara
grafton@psych.ucsb.edu
Kent A. Kiehl
Department Psychology, University of New Mexico
kkiehl@mrn.org

Abstract

Non-negative matrix factorization (NMF) is an excellent tool for unsupervised parts-based learning, but proves to be ineffective when parts of a whole follow a specific pattern. Analyzing such local changes is particularly important when studying anatomical transformations. Hence, we propose a supervised method that incorporates a regression constraint into the NMF framework and learns maximally changing parts in the basis images, called Regression based NMF (RNMF). The algorithm is made robust against outliers by learning the distribution of the input manifold space, where the data resides. Two of our main goals are to achieve good local region visualization as well as recognition accuracy. By incorporating a gradient smoothing and independence constraint into the factorized bases, visual appeasement and accuracy are accomplished. We apply our technique to a synthetic dataset and structural MRI brain images of subjects with varying ages. We find that the localized regions which are expected to be highly changing over age are manifested in our significant basis and we also achieve the best performance compared to other statistical regression and dimensionality reduction techniques.

1. Introduction

Magnetic Resonance imaging (MRI) is a popular imaging modality used to study the anatomy of the brain. The advancement of MRI image acquisition quality has led to the development of many automated and computer-assisted methods in the field of medical image analysis. Population based inference of medical images, such as normative anatomy has become an important image analysis problem. For example, understanding normal changes of anatomy as a function of age is important for distinguishing changes due to disease progression. A central challenge is to develop algorithms that can characterize individual data based on data derived from cross-sectional samples. Consider the data set in Figure 1: Can we predict the age of individuals given their anatomical brain scan? Such a task can be approached by previous a priori knowledge of region of interest (ROI), followed by extracting statistical information from these specific ROIs. However, these methods are highly laborious and require a lot of manual intervention. Moreover, it is not always known in advance which regions certain diseases might affect. To overcome these limitations of ROI analyses, alternative approaches based on voxel-wise analysis without a priori hypothesis have been developed in the past several years, but a fundamental limitation of such approaches is that they cannot identify subtle differences and also tend to be computationally expensive [1, 7].

To develop an accurate predictor of a dependant variable (age) from a set of cross-sectional data, two issues need to be addressed. First, we need a tool to extract the most relevant latent regions from the high dimensional image space. Second, this information needs to be supplied to a pattern recognition tool that provides information regarding the trend of the region across the individual brains. In this paper we propose a matrix factorization method to simultaneously perform dimensionality reduction and deduce intrinsic features that regress, as well as quantize their transformation, thus dealing with the above mentioned problems in a unified and principled manner.

There are many well established matrix factorization algortihms, such as Principal Component Analysis (PCA) [2,

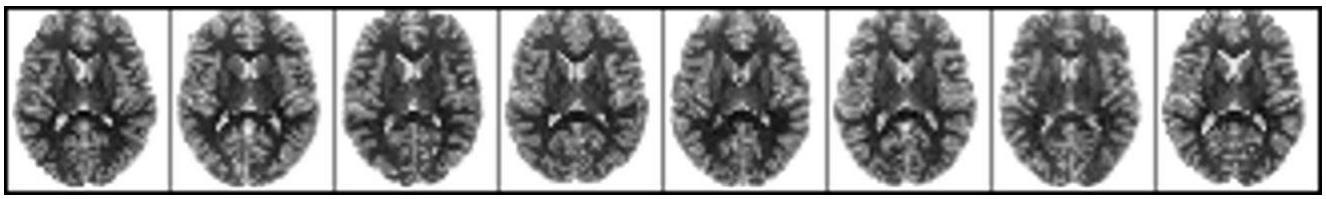


Figure 1. A mid-axial slice is presented for a sample of images used. The images ordered according to increasing patient age from 14 (left) to 40 (right)

14, 27], Independent Component Analysis (ICA) [2, 3, 5], and Non-negative matrix factorization (NMF) [16, 17], that learn to represent the data as a linear combination of basis images; however, each algorithm factorizes the input into these basis vectors subject to different constraints.

Our approach uses NMF [16], an algorithm that learns a parts-based representation of the data. Previous studies have shown that there is physiological and psychological evidence for parts based representation in the human brain [23] and hence NMF has received considerable attention in recent years. It imposes non-negativity constraints on the factored matrices, thus allowing only additive, not subtractive, combinations of the basis vectors, compatible with the intuitive notion of combining parts from a whole.

Our focus is to estimate the age given an individual's structural MRI brain scan. Furthermore, we want to identify and visualize the transforming region of interest (ROI) and understand the evolution of the ROI as age varies. As we want to capture the regressing ROI, we use NMF in conjunction with a regression constraint to identify the trend of the ROI. As shown in Figure 2, the regression curve shows an average increasing trend which represents the expansion of the lateral ventricles as the age increases, which is well known in the medical literature of aging [10, 21, 22]. Using our proposed method we are able to isolate the lateral ventricle region in the brain which shows maximal change and depict its expansion with respect to age.

In this paper, we propose a novel supervised NMF algorithm called regression based non-negative matrix factorization (RNMF). Our contributions are as follows:

- A new supervised NMF model, that captures local parts representing the variability evident in the data by exploiting labeled information of the data. This modification directs the encoding coefficients in the factorization to reflect the trend within these features.
- Improving localization of the basis images by imposing a novel gradient based smoothing constraint which also enhances visual appeal. The convergence proof with this constraint is elaborated in Appendix A.
- By learning the distribution of the data in the input manifold space, we strengthen the robustness of the algorithm to outliers.

The rest of the paper is organized as follows: Section 2.1 introduces the original NMF algorithm. This is followed by a brief overview of regression in Section 2.2 and discussion of learning the manifold space in Section 2.3. We address our proposed algorithm in the subsequent Section 3. Experiments and results with synthetic and structural brain MRI data-sets with RNMF are illustrated in Section 4.

2. Review

2.1. Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF) [16, 17], unlike other methods such as Principal Component Analysis (PCA) [14], is distinguished by its use of non-negativity constraints. NMF has attracted considerable attention [15, 13, 20, 4, 24, 12, 29] because of its many advantages, such as simple yet efficient decomposition of the data, definite physical meaning to parts without negative values, and its lower storage requirement.

Similar to SVD and PCA, NMF decomposes a set of high dimensional vectors to representative lower dimensional vectors using a set of bases under certain constraints. NMF decomposes a non-negative matrix $(\mathbf{V})$ to a set of non-negative basis $(\mathbf{W})$ and corresponding non-negative coefficients $(\mathbf{H})$,

$$\mathbf{V}_{n\times n_t} \approx \mathbf{W}_{n\times m}\mathbf{H}_{m\times n_t} \tag{1}$$

where $\mathbf{V} = [v_{i,j}] = [\mathbf{v}_1, ..., \mathbf{v}_{nt}]$ is a n x n_t matrix , n is the total number of pixels in each image, $\mathbf{v}_j$ is the jth input image represented as a column vector, and n_t is the number of training images. We denote the basis matrix $\mathbf{W} = [w_{i,j}] = [\mathbf{w}_1, ..., \mathbf{w}_m]$ as an n x m matrix (where $(n + n_t)m < nn_t$). The low dimensional embedding of every column of $\mathbf{V}$ is the corresponding column in $\mathbf{H} = [h_{i,j}] = [\mathbf{h}_1, ..., \mathbf{h}_{nt}]$. This factorization is achieved by minimizing the divergence between $\mathbf{V}$ and $\mathbf{WH}$ with the constraints that both should be non-negative. The divergence between $\mathbf{V}$ and $\mathbf{WH}$ is defined as [16, 17]

$$D(\mathbf{V}||\mathbf{WH}) = \sum_{i,j}\left(v_{ij}\log\frac{v_{ij}}{y_{ij}} - v_{ij} + y_{ij}\right), \tag{2}$$

where $\mathbf{WH} = \mathbf{Y} = [y_{ij}]$.

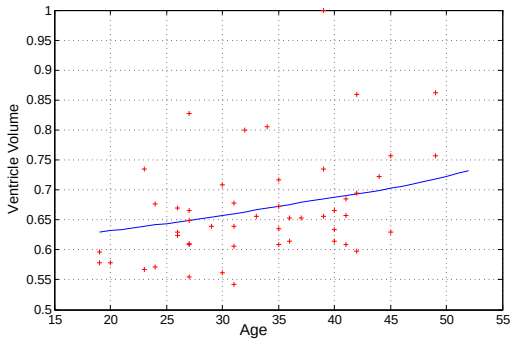


Figure 2. Kernel Regression on ventricle volume versus patient age.

NMF tries to mimic the way humans perceive visual information as a composite of simpler objects. Its basis vectors contain localized features that correspond better with intuitive notions of the parts of the images. While eigenfaces have a statistical interpretation as the directions of largest variance, many of them do not have obvious visual interpretation. However, the additive parts learned by NMF are not necessarily localized and intuitive. In addition, our main goal is to identify transforming regions which NMF fails to do.

2.2. Regression

Regression [11] is a technique used to estimate the relationship, on average, between independent random variables and dependent random variables. A large body of techniques for carrying out regression analysis has been developed and widely used for prediction of dependent variable. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression, such as kernel regression, is also a popular methodology where the regression function can lie in a specified set of functions (kernels). In Figure 2 above, we illustrate the use of kernel regression to demonstrate the effect of age on ventricle volume in the brain.

In subspace analysis, there have been several techniques to simultaneously perform dimensionality reduction and regression [18, 19]. These algorithms reduce the data to a subspace whereafter simple regression techniques can be used for prediction. However, the bases computed by these methods do not highlight the parts that regress, and lack visual appeal. Support vector regression (SVR) [28] is an approach that estimates the regression function directly in the higher dimension. However, SVR fails to indicate the varying region and its corresponding trend. Our proposed algorithm isolates the maximally regressing part and quantifies the change which these algorithms fail to indicate.

2.3. Learning the distribution in Manifold Space

Images can be considered as points in a high dimensional space. We want to learn how these points are distributed on a high dimensional manifold. Assuming a smooth manifold, local neighborhood geodesic distance can be approximated by Euclidean distance. However, for distant points in the high dimensional space the approximation is not valid. Thus, similar to the manifold learning algorithm Isomap [26], a k-nearest neighbor graph with the Euclidean distance as edge weights is constructed; after which the shortest path distance matrix $\mathbf{D} = [d(\mathbf{v}_i, \mathbf{v}_j)]$ on the graph is calculated using Floyd's algorithm [6].

Understanding the distribution of the data based on image content, irrespective of the labels, will help avoid the influence of outliers. Section 3 details how this is imposed as a constraint to enhance the performance of our algorithm.

3. Regression based Non-Negative Matrix Factorization (RNMF)

The main goal of this paper is to simultaneously perform dimensionality reduction and capture regressing features within a cross-sectional data-set. Standard NMF on its own fails to realize the underlying intrinsic regressing parts within the data, which is essential to real world applications. We therefore incorporate a novel regression constraint within the factorization framework to achieve dimensionality reduction and emphasize local features which regress. Furthermore, the robustness of our algorithm is enhanced by learning the distribution of the data in the higher dimensional manifold space. In addition, for more desirable feature visualization purposes, we integrate our method with a gradient based smoothing constraint. This also helps to obtain highly localized contiguous parts. We describe our approach in detail in the following section.

3.1. Constraints

The NMF model defined by the constrained minimization of Equation 2 does not impose any constraints given the labels of the data, being an unsupervised learning technique. We propose a supervised factorization algorithm with the following constraints to achieve our goal:

1. In order to produce regressing features, we want the data to be smoothly separated; which implies, if labels y_i, y_j for corresponding data points $\mathbf{v}_i$, $\mathbf{v}_j$ are close, then the lower dimensional embedding $\mathbf{h}_i$, $\mathbf{h}_j$ should reflect this proximity. Hence, under this assumption we incorporate the regression methodology by imposing the following:

$$S_R = \min \frac{\sum_{i,j} f(y_i, y_j)(\mathbf{h}_i - \mathbf{h}_j)^T(\mathbf{h}_i - \mathbf{h}_j)}{\sum_{i,j} f(y_i, y_j)}. \quad (3)$$

The function $f(\cdot)$ is a weighting function with all positive values. We define $f(y_i, y_j) = \exp(\frac{-|y_i - y_j|}{t})$ (heat kernel), and t depends on the range of the labels, so as the distance between data increases in the dependent variable space, the weight decreases accordingly.

We could have segmented the data-set into several classes based on some fixed boundaries and applied the Fisher constraint similar to [29]. However, the results would have been highly dependent on how the data was sectioned and the number of classes used. Moreover, this approach would fail to take into account similarities between the different classes. Hence, we use the heat kernel to account for the continuous nature of the target variable.

The above constraint only considers the labels assigned to each sample in the cross-sectional data. However, it is also necessary to ensure some robustness to outliers in the data-set. For example, if a pair of points have similar labels, but are highly distant in the input space, we would want to reduce the influence of such a pair on the overall system. Hence, we learn the distribution of the data in the higher manifold space by calculating the geodesic distances between each individual sample as mentioned in Section 2.3. We then compute weights that are inversely proportional to the geodesic distance between the sample points, thus penalizing outliers. This is imposed by

$$m_{ij} = \frac{\min\limits_{i \neq j} d(\mathbf{v}_i, \mathbf{v}_j)}{d(\mathbf{v}_i, \mathbf{v}_j)} f(y_i, y_j), \tag{4}$$

where $d(\mathbf{v}_i, \mathbf{v}_j)$ is calculated as mentioned in Section 2.3. Including this in the first constraint above we get,

$$S_{Rm} = \min \frac{\sum_{i,j} m_{ij} (\mathbf{h}_i - \mathbf{h}_j)^T (\mathbf{h}_i - \mathbf{h}_j)}{\sum_{i,j} m_{ij}}. \tag{5}$$

2. In order to highlight and segregate the most regressing feature in the data-set, good visual perception as well as localization of the bases are necessary. The aim is to compute basis images ($\mathbf{W_j}$) that are less noisy and restricted to comprise of compact regions. Furthermore, we want to ensure that local parts are captured as contiguous regions and not separated among different basis. To accomplish this, we include a novel gradient based smoothing constraint, where we minimize the energy of the gradient of each of the basis $\mathbf{W_j}$ in the image co-ordinates. This is achieved by enforcing the following constraint:

$$S_G = \min \sum_j \int |\nabla \mathbf{W}_j(\mathbf{x})|^2 d\mathbf{x}. \tag{6}$$

3. Similar to LNMF [20] we impose $\sum_i u_{ii}$ = min, where $\mathbf{U} = [u_{ij}] = \mathbf{W}^T\mathbf{W}$. This constraint attempts to minimize the number of basis components required to represent $\mathbf{V}$. This implies that a basis vector should not be further decomposed into more components. Furthermore, to reduce redundancy between different bases, LNMF [20] attempts to make bases as orthogonal as possible. This is imposed by $\sum_{i \neq j} u_{ij}$ = min. Therefore

$$S_O = \ min \sum_{i,j} u_{ij}. \tag{7}$$

The inclusion of the above three constraints leads to the following constrained divergence to be minimized:

$$D(\mathbf{V}||\mathbf{WH}) = \min \Bigg[\sum_{i,j} \left(v_{ij} \log \frac{v_{ij}}{y_{ij}} - v_{ij} + y_{ij} \right) + \alpha S_{Rm} + \beta S_G + \gamma S_O \Bigg] \tag{8}$$

where $\mathbf{WH} = \mathbf{Y} = [y_{ij}]$, and $\alpha, \beta, \gamma > 0$ are constants. The following update rules can be found by minimization of the above constrained cost function:

$$h_{kl} \leftarrow \frac{-b + \sqrt{b^2 + \left(\sum_i v_{il} \frac{w_{ik} h'_{kl}}{\sum_{k'} w_{ik'} h'_{k'l}} \right) \left(\frac{16 \alpha m_{kl}}{\sum_{i,j} m_{ij}} \right)}}{\left(\frac{8 \alpha m_{kl}}{\sum_{i,j} m_{ij}} \right)} \tag{9}$$

$$w_{kl} \leftarrow \frac{w_{kl} \sum_j v_{kj} \frac{h_{lj}}{\sum_{k'} w_{k'l} h_{lj}}}{\sum_j h_{lj} + \beta g_{kl} + \gamma \sum_j w_{kj}} \tag{10}$$

$$w_{kl} \leftarrow \frac{w_{kl}}{\sum_{k'} w_{k'l}} \tag{11}$$

where,

$$b = 1 - \frac{4\alpha}{\sum_{i,j} m_{ij}} \sum_{k'} (h_{kk'} m_{k'l}) \tag{12}$$

and $[g_{ij}] = [\mathbf{g}_1, \mathbf{g}_2, ...\mathbf{g}_n]$ where $-\nabla^2 \mathbf{W}_j$ (in the image coordinates) is vectorized as $\mathbf{g_j}$. The proof of convergence is provided in Appendix A.

4. Experiments and Results

We present the performance of our proposed algorithm with two data-sets: synthetic data and structural MRI brain scans of subjects with varying ages. We also compare our algorithm against two statistical regression methods, Sliced Inverse Regression (SIR) [18] and Principal Hessian Directions (PHD) [19], as well as PCA [2, 14, 27], SVR [28], LNMF [20] and the original NMF [16, 17] techniques.

Figure 3. Example of synthetic data for I_j where j = 1, 50, 100, 150, 200, $\frac{\sigma}{\eta}$ = 0.14. p_1, p_2, p_3, p_4 corresponds to bottom right (yellow), bottom left (green), top left (pink), top right (purple) blocks respectively. This figure is best viewed in color.

4.1. Synthetic Data

We begin our analysis on a synthetic data-set as illustrated in Figure 3. The data-set consists of 200 images $[I_j]_{j=1,...,200}$, where each image is of size 200×200. Every image consists of four distinct square parts p_i, $i \in [1, 2, 3, 4]$, locally varying in intensity with a specific regressing pattern. The data-set is labeled using $y_j = j$. The following are the variations formulated for each square box

$$I_j(\mathbf{x}) = \frac{\eta}{200} j + N, \mathbf{x} \in p_1 \tag{13}$$

$$I_j(\mathbf{x}) = \frac{\eta}{200}(200 - j) + N, \mathbf{x} \in p_2 \tag{14}$$

$$I_j(\mathbf{x}) = \frac{\eta}{200}(100 - |j - 100|) + N, \mathbf{x} \in p_3 \tag{15}$$

$$I_j(\mathbf{x}) = \frac{\eta}{200}(|j - 100|) + N, \mathbf{x} \in p_4, \tag{16}$$

where η is an arbitrary constant that models the foreground intensity and $N = \mathcal{N}(\mu,\sigma^2)$. We randomly select 100 images for training and the rest 100 for testing. We create 5 synthetic datasets where η = constant, $\mu = 0$ and vary σ between 5-25. Hence, we have data-sets with varying noise levels and divided into the same training and testing sets for our analysis.

Results : The problem in hand is to predict the dependent variable (label) y_j given the independent variable (image) I_j. In addition, good visualization of the identified regressing regions and computing the associated change is necessary. We find that RNMF significantly performs better than other methods in prediction. In addition, in Figure 5(b) we compare the error (mean of the difference between predicted and actual value) vs increasing normalized noise level ($\frac{\sigma}{\eta}$) and observe RNMF outperform all the other methods. Figure 4 shows the ability of our method to factor out different regressing parts in separate bases, consistent with the parts based notation in Equations 13-16. Other methods fail to decompose the images into their constituent parts. Also, the coefficients obtained by RNMF (Figure 4(a)) reflect the nature of the change unlike other methods. We intuitively expect RNMF to perform better than NMF, LNMF and PCA as these are unsupervised learners. RNMF being a parts-based factorization, captures local parts efficiently, thus outperforming global regression techniques such as SIR, PHD and SVR. We also note that irrespective of the choice of the number of bases, the recognition ability of RNMF supasses other techniques and it remains consistent as shown in Figure 5(a).

Methods	Number of basis					
	5	10	15	20	25	30
PCA	8.1	**7.1**	11.5	13.0	9.6	9.9
SIR	**6.4**	8.1	8.1	8.1	8.1	8.1
PHD	9.2	**6.1**	6.7	6.5	6.9	7.3
SVR	9.7					
LNMF	8.0	8.5	6.9	**6.3**	7.4	7.1
NMF	**6.9**	7.5	7.1	7.5	7.4	9.9
RNMF	5.7	6	5.2	**4.7**	5.2	5.5

Table 1. Average recognition error with varying the number of bases. Bold numbers represent the least recognition error for each method.

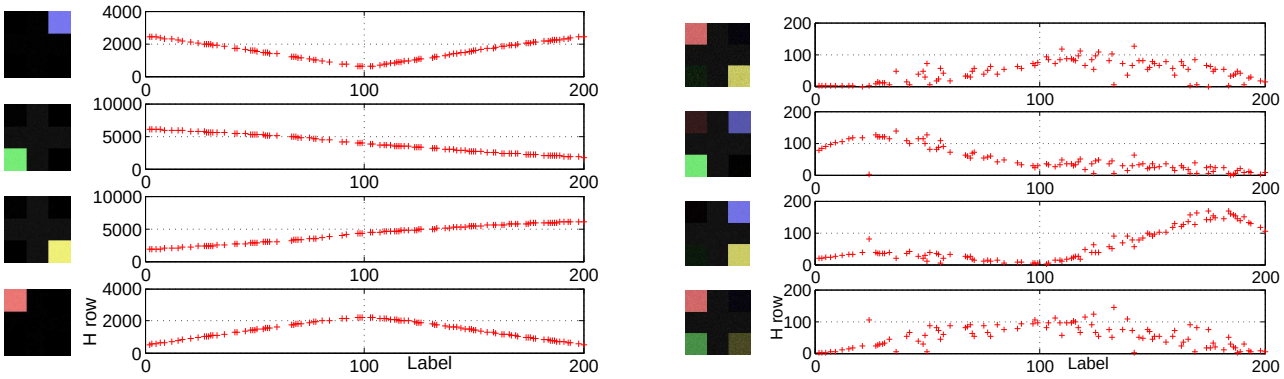
4.2. Structural MRI Brain Data

The data-set consists of T1-MPRAGE structural MRI scans of 200 incarcerated prisoners acquired at a spatial resolution of 1mm × 1mm × 1mm using a 1.5 Tesla head-only scanner. We preprocess the images, first skull extraction followed by spatial alignment to an atlas using affine registration using the standard software program FMRIB Software Library (FSL) [25, 30]. All brains are normalized, where each voxel contains a value between 0 and 1. Each of our images are 40×45×40 in size after downsampling them by 4. Each scan has an associated age (y_j) that ranges from 14-58 years.

The scans are randomly partitioned into a training subset of 130 and a test set of 70 images. The training set is then used to learn the regressing basis components and the test set for evaluation. All the compared methods use the same training and testing set. The constants α, β and γ are empirically determined such that we get non-negative updates.

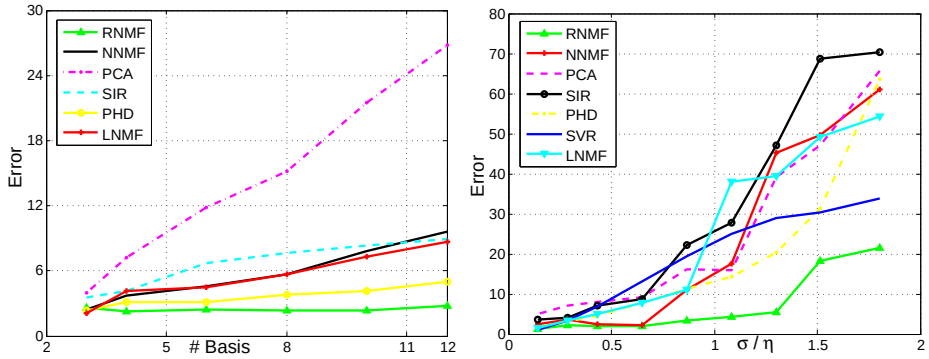
Results : Our focus is to estimate the age given an individual's structural MRI brain scan. Furthermore, we want to identify and visualize the maximally transforming anatomical part and understand its evolution as age varies. We compare our method with other techniques mentioned in Section 4. We noticed that the bases for NMF, PCA, PHD and SIR were highly holistic. Moreover, the encoding coefficients of LNMF and NMF hardly convey information regarding the regression pattern (6(e), 6(f)). Figure 6 shows the most significant basis for RNMF, LNMF and NMF. Significance of bases was computed according to the regression coefficient, defined as the slope of the best fit line to the distribution of the corresponding row in H. Also SVR, PHD and SIR give global bases where the regressing part cannot be localized and segregated easily.

Our RNMF method leads to a parts-based representation, where solely the expanding lateral ventricle is captured in



(a) (b)

Figure 4. Comparing basis and encoding coefficients of RNMF and NMF for $\frac{\sigma}{\eta}$=0.14 . Figure best viewed in color.



(a) Performance based on changing number of basis (b) Performance based on changing noise level

Figure 5. Comparing recognition error of the different methods with regards to changing noise level and number of bases. Figure best viewed in color.

the basis component. The corresponding encoding coefficients 6(d) reflect the expansion of the region. Figure 6(a) shows the horizontal slices of the brain captured in the basis depicting the lateral ventricle region. Furthermore, as the number of bases increases, Figure 6(a) remains to be the most significant basis using RNMF, unlike other methods. Note that our findings in Figure 6(d) agree with volume-based regression analysis in Figure 2 and are consistent with the results obtained by [8] and in the medical literature [10, 21, 22], where we see that as age increases the ventricle volume increases. Table 1 shows the recognition errors for the different methods across various numbers of basis components. The output is predicted using the standard multivariate kernel regression on the lower dimensional representation ($\mathbf{H}$). We computed the recognition error as the mean of the absolute age difference between the actual and predicted ages. RNMF achieves the best performance where the mean absolute error is 4.7 years as seen.

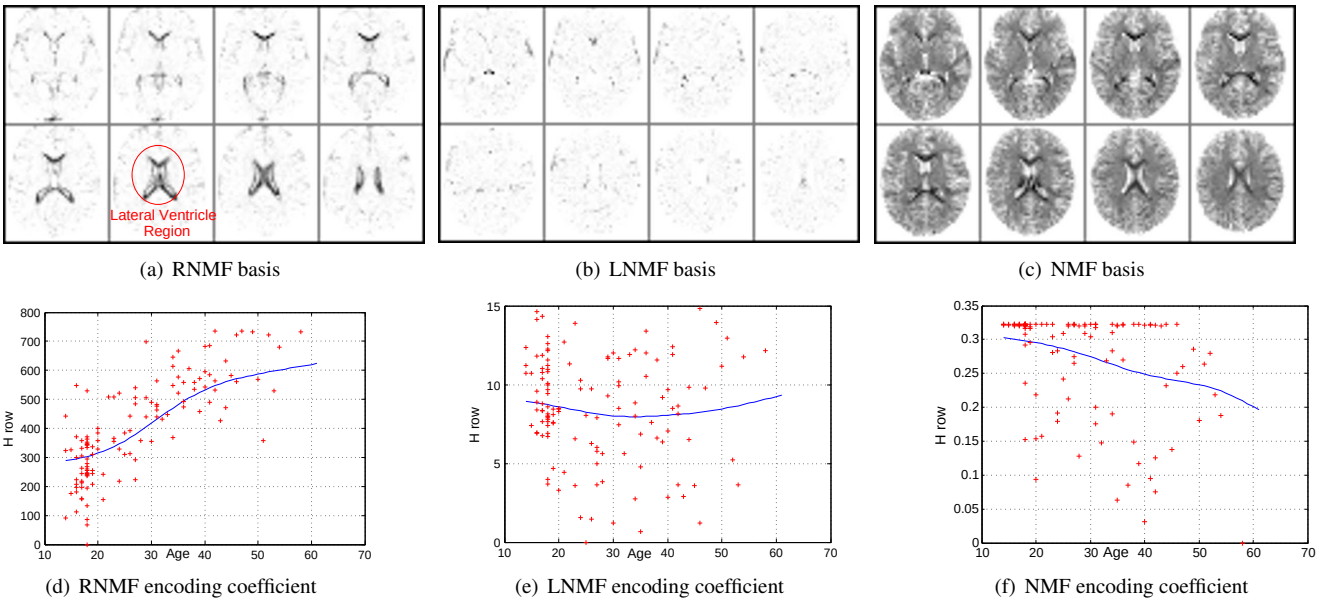
5. Conclusion and Future Work

We have presented a novel method of factorization for regression problems. RNMF bases are more suitable to capture local regressing features. The gradient smoothing constraint improves visual appeal and manifold learning makes the algorithm robust against outliers. Experimental results on synthetic and structural brain MRI data show superior performance in terms of accuracy and visual perception.

We are able to isolate the lateral ventricle region in the brain which shows maximal change with respect to age without prior ROI assumptions. The method is also general in nature and it finds numerous applications in identifying diseased regions which are visually undetectable. Our current work involves identifying the maximally distinguishing anatomical region that can classify psychopaths from non-psychopaths, based on a psychopathy score assigned to them. In addition, we would like to extend our method to a multivariate scenario by modeling the dependence over different output variables, such as age, psychopathy score, etc.

6. Acknowledgments

We would like to thank Dr. Sarang Joshi, Dr. Luca Bertelli, Mehmet Emre Sargin and David Clewett for their help and support throughout the project. We gratefully ac-



(a) RNMF basis (b) LNMF basis (c) NMF basis

(d) RNMF encoding coefficient (e) LNMF encoding coefficient (f) NMF encoding coefficient

Figure 6. Comparing most significant basis and corresponding encoding coefficients of RNMF, LNMF and NMF where number of basis = 20. Shown are the horizontal cross-sections. The highlighted region is the lateral ventrical the maximally changing region with respect to age. As we can see RNMF explicitly isolates the ventricle region.

knowledge our funding sources including MacArthur Foundation and Public Health Service grant and NSF awards ITR-0331697 and III-0808772.

A. Convergence Proof of RNMF

Our update rules are based on a technique which minimizes an objective function $L(\mathbf{X})$ by using an auxiliary function. $G(\mathbf{X}, \mathbf{X}')$ is defined as an auxiliary function for $L(\mathbf{X})$ if $G(\mathbf{X}, \mathbf{X}') \geq L(\mathbf{X})$ and $G(\mathbf{X}, \mathbf{X}) = L(\mathbf{X})$ are satisfied [9]. If $\mathbf{G}$ is an auxiliary function, then $L(\mathbf{X})$ is non-increasing when $\mathbf{X}$ is updated by

$$\mathbf{X}^{(t+1)} = \arg\min_{\mathbf{X}} G(\mathbf{X}, \mathbf{X}^{(t)}) \tag{17}$$

$$L(\mathbf{X}^{(t+1)}) \leq G(\mathbf{X}^{(t+1)}, \mathbf{X}^{(t)}) \leq G(\mathbf{X}^{(t)}, \mathbf{X}^{(t)}) = L(\mathbf{X}^{(t)})$$

$\mathbf{W}$ is updated by minimizing $L(\mathbf{W}) = D(\mathbf{V}||\mathbf{WH})$ with $\mathbf{H}$ fixed. We construct an auxiliary function for $L(\mathbf{W})$ as

$$G(\mathbf{W}, \mathbf{W}') = \sum_{i,j} v_{ij} \log v_{ij} + \sum_{i,j,k} v_{ij} \frac{w'_{ik} h_{kj}}{\sum_k w'_{ik} h_{kj}} \left[\log(w_{ik} h_{kj}) - \log \left(\frac{w'_{ik} h_{kj}}{\sum_k w'_{ik} h_{kj}} \right) \right] - \sum_{i,j} v_{ij} + \sum_{i,j} (wh)_{ij} + \alpha S_{Rm} + \beta S_G + \gamma S_O. \tag{18}$$

It is easy to verify that $G(\mathbf{W}, \mathbf{W}) = L(\mathbf{W})$, so we will just prove that $G(\mathbf{W}, \mathbf{W}') \geq L(\mathbf{W})$ as follows. We know that $\log(\sum_k w_{ik} h_{kj})$ is a convex function, and the following holds for all i,j and $\sum_k \sigma_{ijk} = 1$:

$$-\log \left(\sum_k w_{ik} h_{kj} \right) \leq -\sum_k \sigma_{ijk} \log \left(\frac{w_{ik} h_{kj}}{\sigma_{ijk}} \right). \tag{19}$$

Let $\sigma_{ijk} = \frac{w'_{ik} h_{kj}}{\sum_k w'_{ik} h_{kj}}$. Then

$$-\log \left(\sum_k w_{ik} h_{kj} \right) \leq -\sum_k \frac{w'_{ik} h_{kj}}{\sum_k w'_{ik} h_{kj}} \left[\log(w_{ik} h_{kj}) - \log \left(\frac{w'_{ik} h_{kj}}{\sum_k w'_{ik} h_{kj}} \right) \right], \tag{20}$$

which is $G(\mathbf{W}, \mathbf{W}') \geq L(\mathbf{W})$.

Now to minimize L($\mathbf{W}$) w.r.t $\mathbf{W}$, we can update $\mathbf{W}$ using $\mathbf{W}^{(t+1)} = \arg\min_{\mathbf{W}} G(\mathbf{W}, \mathbf{W}^{(t)})$ such that $\mathbf{W}$ can be found out by letting $\frac{\partial G(\mathbf{W}, \mathbf{W}')}{\partial w_{kl}} = 0$ for all kl. We make use of the fact $\frac{\partial S_G}{\partial \mathbf{W}_j} = -\nabla^2 \mathbf{W}_j$ (using Variational Calculus) and we get the update rule for $\mathbf{W}$ as

$$w_{kl} \leftarrow \frac{w_{kl} \sum_j v_{kj} \frac{h_{lj}}{\sum_{k'} w_{k'l} h_{lj}}}{\sum_j h_{lj} + \beta g_{kl} + \gamma \sum_j w_{kj}}, \tag{21}$$

where $[g_{ij}] = [\mathbf{g}_1\ \mathbf{g}_2 .... \mathbf{g}_n]$ and $-\nabla^2 \mathbf{W}_j$ (in the image coordinates) is vectorized as $\mathbf{g_j}$. Similarly $\mathbf{H}$ is updated by minimizing $L(\mathbf{H}) = D(\mathbf{V}||\mathbf{WH})$ with $\mathbf{W}$ fixed. Here

$G(\mathbf{H}, \mathbf{H}')$ is constructed as

$$G(\mathbf{H}, \mathbf{H}') = \sum_{i,j} v_{ij} \log v_{ij} + \sum_{i,j,k} v_{ij} \frac{w_{ik} h'_{kj}}{\sum_k w_{ik} h'_{kj}} \left[\log(w_{ik} h_{kj}) - \log \left(\frac{w_{ik} h'_{kj}}{\sum_k w_{ik} h'_{kj}} \right) \right] - \sum_{i,j} v_{ij} + \sum_{i,j} (wh)_{ij} + \alpha S_{Rm} + \beta S_G + \gamma S_O \quad (22)$$

After setting $\frac{\partial G(\mathbf{H}, \mathbf{H}')}{\partial h_{kl}} = 0$ for all kl, we get the following update rule

$$h_{kl} \leftarrow \frac{-b + \sqrt{b^2 + \left(\sum_i v_{il} \frac{w_{ik} h'_{kl}}{\sum_{k'} w_{ik'} h'_{k'l}} \right) \left(\frac{16 \alpha m_{kl}}{\sum_{i,j} m_{ij}} \right)}}{\left(\frac{8 \alpha m_{kl}}{\sum_{i,j} m_{ij}} \right)} \quad (23)$$

where,

$$b = 1 - \frac{4\alpha}{\sum_{i,j} m_{ij}} \sum_{k'} (h_{kk'} m_{k'l}). \quad (24)$$

References

[1] J. Ashburner and K. Friston. Voxel-based morphometry-the methods. *Neuroimage*, 11(6):805–821, 2000.

[2] K. Baek et al. PCA vs. ICA: A comparison on the FERET data set. In *Joint Conference on Information Sciences, Durham, NC*, pages 824–827. Citeseer, 2002.

[3] A. Bell and T. Sejnowski. The independent components of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997.

[4] X. Chen, L. Gu, S. Li, and H. Zhang. Learning representative local features for face detection, IEEE. In *Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1126–1131, 2001.

[5] P. Comon et al. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[6] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to algorithms*. The MIT press, 2001.

[7] C. Davatzikos, A. Genc, D. Xu, and S. Resnick. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. *NeuroImage*, 14(6):1361–1369, 2001.

[8] B. Davis, P. Fletcher, E. Bullitt, and S. Joshi. Population shape regression from random design data. In *Proceeding of ICCV*, 2007.

[9] A. Dempster et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[10] Guttmann et al. White matter changes with normal aging. *Neurology*, 50(4):972, 1998.

[11] W. Hardle. *Applied nonparametric regression*. Cambridge University Press Cambridge, 1990.

[12] M. Heiler and C. Schnorr. Learning sparse representations by non-negative matrix factorization and sequential cone programming. *The Journal of Machine Learning Research*, 7:1407, 2006.

[13] P. Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469, 2004.

[14] I. Jolliffe. *Principal component analysis*. 2002.

[15] K. Kochi, D. Lehmann, R. Pascual-Marqui, et al. Nonsmooth Nonnegative Matrix Factorization (nsNMF). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):415, 2006.

[16] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[17] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, pages 556–562, 2001.

[18] K. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, pages 316–327, 1991.

[19] K. Li. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992.

[20] S. Li, X. Hou, H. Zhang, and Q. Cheng. Learning spatially localized, parts-based representation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1. IEEE Computer Society, 2001.

[21] M. Matsumae, R. Kikinis, et al. Age-related changes in intracranial compartment volumes in normal adults assessed by magnetic resonance imaging. *Journal of neurosurgery*, 84(6):982–991, 1996.

[22] Z. Mortamet, B et al. Effects of healthy aging measured by intracranial compartment volumes using a designed mr brain database. *LECTURE NOTES IN COMPUTER SCIENCE*, 3749:383, 2005.

[23] S. Palmer. Hierarchical structure in perceptual representation. *Cognitive Psychology*, 9(4):441–474, 1977.

[24] R. Sandler and M. Lindenbaum. Nonnegative Matrix Factorization with Earth Movers Distance Metric. CVPR, 2009.

[25] S. Smith, M. Jenkinson, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23:208–219, 2004.

[26] J. Tenenbaum, V. Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[27] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[28] V. Vapnik. *The nature of statistical learning theory*. Springer Verlag, 2000.

[29] Y. Wang et al. Fisher non-negative matrix factorization for learning local features. In *Proc. Asian Conf. on Comp. Vision*, pages 27–30. Citeseer, 2004.

[30] M. Woolrich et al. Bayesian analysis of neuroimaging data in FSL. *NeuroImage*, 45(1S1):173–186, 2009.