A General and Unifying Framework for Feature Construction, in Image-Based Pattern Classification

Nematollah Batmanghelich[1], Ben Taskar[2], and Christos Davatzikos[1]

[1] Section of Biomedical Image Analysis, Raddiology Department, University of Pennsylvania, Philadelphia PA 19014, USA
[2] Computer and Information Department, University of Pennsylvania, Philadelphia PA 19104, USA
{batmangh@seas,taskar@cis,christos@rad}.upenn.edu

Abstract. This paper presents a general and unifying optimization framework for the problem of feature extraction and reduction for high-dimensional pattern classification of medical images. Feature extraction is often an ad hoc and case-specific task. Herein, we formulate it as a problem of sparse decomposition of images into a basis that is desired to possess several properties: 1) Sparsity and local spatial support, which usually provides good generalization ability on new samples, and lends itself to anatomically intuitive interpretations; 2) good discrimination ability, so that projection of images onto the optimal basis yields discriminant features to be used in a machine learning paradigm; 3) spatial smoothness and contiguity of the estimated basis functions. Our method yields a parts-based representation, which warranties that the image is decomposed into a number of positive regional projections. A non-negative matrix factorization scheme is used, and a numerical solution with proven convergence is used for solution. Results in classification of Alzheimers patients from the ADNI study are presented.

1 Introduction

Voxel-based analysis (VBA) has been widely used in the medical imaging community. It typically consists of mapping image data to a standard template space, and then applying voxel-wise linear statistical tests on a Jacobian determinant [1], [2], transformation-residuals [3], or tissue density maps [4], [5] or directly on voxel intensity (e.g. diffusion imaging [6]). It therefore identifies regions in which two groups differ (e.g. patients and controls [2]), or regions in which other variables (e.g. disease severity [7]) correlate with imaging measurements. However, this method has limited ability to identify complex population differences, because it does not take into account the multivariate relationships in data [8], [9]. Moreover, since typically no single anatomical region offers sufficient sensitivity and specificity in identifying pathologies that span multiple anatomical regions, it has very limited diagnostic power on an individual basis. In other words, values of voxels or ROIs showing significant group difference are not necessarily good discriminants when one wants to classify individuals into groups.

J.L. Prince, D.L. Pham, and K.J. Myers (Eds.): IPMI 2009, LNCS 5636, pp. 423–434, 2009.

In order to overcome the limitations, high-dimensional pattern classification methods have been proposed in the relatively recent literature [9,10,11,12,13], which capture multi-variate nonlinear relationships in the data, and aim to achieve high classification accuracy of individual scans. A fundamental difficulty in these methods has been the availability of enough training samples, relative to the high dimensionality of the data. A critical problem has therefore persisted in these methods, namely how to optimally perform feature extraction and selection, i.e. to find a parsimonious set of image features that best differentiate between two or more groups, and which generalize well to new samples.

Feature reduction methods can be categorized into two general families: 1) feature selection and 2) feature construction [14]. Feature selection methods (e.g. SVM-RFE [15]) have two problems: first, they do not scale up for medical images; second, they do not consider domain knowledge (in our case: the fact that data is coming from images) thus they may end up selecting a subset of features which is not biologically interpretable. Another family of feature reduction methods includes feature construction like PCA, LDA or other linear or nonlinear transformations. These methods can take into account domain knowledge but they are challenged by two issues: first, constructed features do not have local support, but are typically extracted from spatially extensive and overlapping regions; moreover, they use both positive and negative weights, which render difficult anatomical interpretability. Finally, the number of basis vectors is usually bounded by the number of samples, which is usually less than the dimensionality of features.

In this paper, we propose a novel method which falls into the feature construction category. Finding optimal linear construction can be viewed as finding a linear transformation, i.e. basis matrix, which is to be estimated from data according to some desired properties that are discussed next. 1) The basis must be biologically meaningful: this means that a constructed basis vector should correspond to contiguous anatomical regions preferably in areas which are biologically related to a pathology of interest. Having local spatial support can be viewed mathematically as sparsity of a basis vector in combining voxel values. 2) The basis must be discriminant: we are interested in finding features, i.e. projection onto the basis, that construct spatial patterns that best differentiate between groups, e.g. patients and controls. 3) The basis must be representative of the data: in order to represent data, we derive a basis matrix with aforementioned properties and corresponding loadings. Matrix factorization has been adopted as a framework. Having simultaneously representative and parsimonious representation of an image is usually referred to parts-based representation in the literature. A specific variant of Matrix Factorization (MF) which is confined to be nonnegative (NMF) has been shown experimentally [16], and under some conditions mathematically [17], to yield parts-based representations of an image. Since general NMF does not consider that underlying data is an image, we have introduced a Markovian prior to address this issue. Furtheremore, we have an extra prior to enforce sparsity (parts-based representation) of an image. 4) Generalization: the proposed method is general and can be applied to a wide

variety of problems and data sets without significant adjustments. In this paper, we have formulated our problem as an optimization problem that seeks to satisfy the four criteria above. Moreover, we proposed a novel numerical solution with a proof of convergence to solve it. Unlike LDA and PCA, the number of basis vectors are not confined with number of samples in our method thus we are able to have more basis vectors than samples.

In the Methods section, we first discuss the idea of matrix factorization in general and NMF in particular (Sect.2.1). In the subsequent sections, a likelihood term (Sect.2.2) and proper regularization terms are introduced (Sect.2.3,2.4). In Sect. 2.5, the final optimization problem is formed and a proper method is suggested to solve it. In the Results section (Sect.3), we apply our method to the problem of classification of Alzheimer's disease patients and healthy controls.

2 Methods

2.1 General Formulation

Let's assume that we collect data into a matrix, $X \in \mathbb{R}^{+D\times N}$, such that each column x_i represents one image. This can be done by lexicographical ordering of voxels. D is number of voxels and N is number of samples. For this case, we assume that x_i's reside in positive quadrant which is a reasonable assumption for medical images. The goal is to decompose data matrix, X, into a positive matrix, B, which is a matrix whose columns are constructed basis vectors, and a loadings matrix , C, which holds corresponding loadings of the basis, namely $X \approx BC$. The elements of C will form the features extracted from the data via projection on B; they will be subsequently used for classification. In the literature, this decomposition is called Non-Negative Matrix Factorization (NMF). It is straightforward to verify that this is an ill-posed problem. Hence, a regularization is necessary. We formulate the problem as a MAP (Maximum a Posteriori) estimation problem as follows:

$$p(B,C|X) = \frac{p(X|B,C)p(B,C)}{p(X)} = \frac{p(X|B,C)p(B)p(C)}{p(X)} \tag{1}$$

Here, we assumed that B and C are independent. Therefore, the MAP estimation problem is formulated as an optimization problem as follows:

$$\max_{B,C} \log p(B,C|X) \equiv \max_{B,C} \log p(X|B,C) + \log p(B) + \log p(C) \tag{2}$$

in which the first term on the right hand side is a likelihood term and the second and third terms are priors for B and C respectively. Thus, we need to choose proper priors and likelihood function according to our problem. In general, NMF can be written as the following optimization problem:

$$\min_{B,C>0} D(X;BC) + \alpha(B) + \beta(C) \tag{3}$$

where $D(X;BC)$ is a negative likelihood function and measures the goodness of fit, and where the second ($\alpha(B)$) and third ($\beta(C)$) terms form negative log priors on B and C. Next, we discuss different choices for $D(.,.)$, $\alpha(.)$, and $\beta(.)$.

2.2 Likelihood Term: $D(X; BC)$

As it is discussed in [18], given a convex function $\varphi : S \subseteq \mathbb{R} \rightarrow \mathbb{R}$, Bregman divergence is a family of $D(.,.)$ functions which are defined as follows $D_\varphi : S \times \text{int}(S) \rightarrow \mathbb{R}_+$:

$$D_\varphi(x; y) := \varphi(x) - \varphi(y) - \varphi'(y)(x - y) \tag{4}$$

where $\text{int}(S)$ is the interior of set S. For cases in which x and y are matrices, it can be augmented as summation over all elements of a matrix:

$$D_\varphi(X; Y) := \sum_{ij} D_\varphi(x_{ij}, y_{ij}) \tag{5}$$

In this paper, we used $\varphi(x) = x \log x$ which readily converts (5) to the KL-Divergence:

$$D_\varphi(X; BC) = \sum_{ij} x_{ij} \log \frac{x_{ij}}{\sum_k b_{ik} c_{kj}} - \sum_{ij} x_{ij} + \sum_{ijk} b_{ik} c_{kj} \tag{6}$$

It is worth mentioning that other choices for φ are also possible (e.g. $\frac{1}{2}x^2$) and they yield other distance measures (e.g. Frobenius distance between matrices).

2.3 Regularizing the Basis: $\alpha(B)$

The regularization term can be broken down into two terms according to respective criteria that will be discussed in more detail in this section:

$$\alpha(B) = \alpha_1(B) + \alpha_2(B) \tag{7}$$

In our implementation, each regularization term has a weighting term which determines its contribution, however, we have omitted the weighting terms for the sake of simplicity in the notation.

It is reasonable to assume that anatomical regions are expected to display similar structural and functional characteristics, hence voxels should be grouped together into regional features. As discussed in the Introduction, local support and sparsity are two desirable properties which both can be achieved using the following terms:

$$\alpha_1(B) = \mathbf{1}^T B^T B \mathbf{1}, \quad \|b_i\|_1 = 1 \tag{8}$$

In order to see why this regularization enforces part-based representation, we should interpret it mathematically. Part-based representation means that we do not want our basis vectors, b_i, to have a lot of overlap with each other. Considering the fact that the basis are positive (hence, bounded below), having the least overlap could be translated to orthogonality. Mathematically speaking, $< b_i, b_j > \approx 0$ if $i \neq j$ which means that off-diagonal elements of $B^T B$ should be minimized.

It is also worth mentioning that it has been shown empirically [16] and under some mild conditions mathematically [17] that NMF yields sparse basis. Nevertheless, equality constraint in (8) in addition to the non-negativity constraint enforces sparsity even further. This ends the justification of the terms introduced in (8).

This is the first criterion for the prior over B and was mentioned earlier in [19]; nevertheless this is not enough when one deals with image data. Diseases typically affect anatomy and function in a somewhat continuous way. Therefore, we would prefer that b_i represents smooth and contiguous anatomical regions. Although smoothing can be applied as post processing after optimization and deriving B, it is preferable to add a smoothness penalty term to the prior of B. Similar to [20], we exploit the widely used Markov Random Field (MRF) model. In this model, voxels within a neighborhood interact with each other and smoothness of an image is modeled as in the Gibbs distribution as follows:

$$p(I) = \frac{1}{Z} \exp\left(-c\alpha_2(B)\right) \quad \Rightarrow -\log p(I) = c\alpha_2(B) - \log Z \tag{9}$$

where I is a vector made by concatenating image voxels (e.g. lexicographically) and Z is a normalization constant called partition function and c is a constant. $\alpha_2(.)$ is a nonlinear energy function measuring non-smoothness of an image. For basis matrix B, we can write $\alpha_2(B)$ as follows:

$$\alpha_2(B) = \sum_{j=1}^{r} \sum_{i=1}^{D} \sum_{l \in U_i} w_{il} \psi(b_{ji} - b_{jl}, \delta) \tag{10}$$

where r is the number of basis vectors and D is dimensionality of the images and U_i is a set containing the neighborhood indices of the i'th voxel and $\psi(., \delta)$ is a potential function and δ is a free parameter and w_{kl} are weighting factors. There are plenty of choices for the potential function. We adopt a simple quadratic function that has all desired properties, including nonnegativity, strictly increasing, unboundedness and more importantly convexity in addition to the fact that it can be simply represented in a matrix form which will help us to derive an appropriate auxiliary function:

$$\psi(x, \delta) = (\frac{x}{\delta})^2 \tag{11}$$

Adding both terms, α_1 and α_2, for basis, total regularization penalty would become:

$$\alpha(B) = \mathbf{1}^T B^T B \mathbf{1} + \sum_{j=1}^{r} \sum_{i=1}^{D} \sum_{l \in U_i} w_{il} \psi(b_{ji} - b_{jl}, \delta) \tag{12}$$

2.4 Regularizing Coefficients: $\beta(C)$

In this section, we will discuss the regularization term for the coefficient matrix. The main goal of these regularization terms is to boost bases that produce

discriminant features, but also are found consistently across all training samples. We decompose the regularization terms for the C matrix into two terms and describe each one in detail:

$$\beta(C) = \beta_1(C) + \beta_2(C) \tag{13}$$

In our implementation, each regularization term has a weighting term which determines its contribution, however, we have omitted the weighting terms for the sake of simplicity in the notation.

Given the basis matrix, B, the coefficient matrix, C represents new features. If the final goal is classification, discriminative features are preferred. Similar to [21], we use Fisher linear discriminative analysis which is the largest generalized eigen value between within- and between- class matrices when c_{ij} coefficients are considered as new features:

$$\begin{array}{l} S_i = \frac{1}{N_i} \sum_{k \in \mathcal{I}_i} (c_k - \bar{c}_i)(c_k - \bar{c}_i)^T \\ S_W = \frac{1}{2}(S_1 + S_2) \qquad\qquad i = 1, 2 \\ S_B = \frac{1}{2}(\bar{c}_1 - \bar{c}_2)(\bar{c}_1 - \bar{c}_2)^T \end{array} \tag{14}$$

where $\mathcal{I}_i$ is a set containing indices of instances in the i'th class and and c_k is k'th column of matrix C and $\bar{c}_i$ is the mean of new features over i'th class ($\bar{c}_i = 1/N_i \sum_{k \in \mathcal{I}_i} c_k$). S_i is the within-class matrix for the i'th class and S_B is the between-class matrix. Here, we have assumed that we have two classes but the formulation can be easily extended. We would like to maximize the largest generalized eigen value between S_B and S_W, however there is no closed form formulation for that. Instead, we use an approximation as follows [21]:

$$p(C) \propto \exp(\beta_1(C)) \propto \frac{\exp(tr(S_B))}{\exp(tr(S_W))} \Rightarrow -\log p(C) \propto -\beta_1(C) \propto tr(S_B) - tr(S_W) \tag{15}$$

Trace of S_W which is summation of eigen values approximately measures how skewed the classes are, and trace of S_B roughly evaluates how far apart the two classes are. Hence, the more separable the classes are, the lower $\beta_1(C)$ is.

The second criterion for the C matrix is to seek bases which carry maximum image energy. Total *activity* of retained components, i.e. total squared projection coefficients summed over all training images, should be maximized [19]. Effectively, this constraint favors bases that represent components that tend to be present in all samples, and therefore reflect anatomically consistent regions that are likely to generate new samples. Energy of each retained basis is measured by the l_2 norm of c_i^T in which c_i^T is the i'th row of matrix C:

$$\beta_2(C) = -\sum_i \|c_i^T\|_2 \tag{16}$$

Adding up $\beta_1(.)$ and $\beta_2(.)$, yields the final regularization term on C matrix:

$$\beta(C) = \beta_1(C) + \beta_2(C) = tr(S_W) - tr(S_B) - \sum_i \|c_i^T\|_2 \tag{17}$$

2.5 Optimization

We have derived all necessary terms and constraints to form an optimization problem. Given the likelihood function, $D(.,.)$ in (6) and equations for regularization functions on B and C, $\alpha(.)$ in (12) and $\beta(.)$ in (17) and corresponding constraints, the optimization problem is as follows:

$$\begin{array}{ll} \min & D_{\varphi}(X; BC) + \alpha(B) + \beta(C) \\ \text{subject to} & \|b_i\|_1 = 1 \\ & [B]_{ij}, [C]_{ij} \geq 0 \end{array} \tag{18}$$

This formulation is not a convex optimization problem. Therefore, we seek a local minimum. A typical strategy to solve this kind of problem is to fix a block of parameters (e.g. C) and optimize other blocks (e.g. B) and alternate until convergence. If C is fixed, optimization over B is a convex problem. Although norm equality constraints for b_i's are not convex constraints in general, they become linear constraints due to the non-negativity of B. However, by fixing B, we do not have a convex optimization problem in C because in (17), $\sum_i \|c_i\|_2$ (a convex term) has to be maximized, not minimized.

Due to the dimensionality of the problem, we use a first order method to solve it. Similar to Lee et al. [16], we prefer a Multiplicative Update (MU). Multiplicative methods have two advantages: first, if initialization starts inside of a feasible set, as long as the current value of a variable is multiplied by a positive value, the new value of that variable is also positive; hence maintaining positivity constraints is trivial. Second, although MU is derived from gradient descent, it has no parameter like the step size of gradient descent. This makes the MU very easy to implement, except one has to make sure that in each iteration the value of cost function decreases. A common approach for optimization in NMF literature is to propose an *auxiliary* function.

Definition 1. *$Z(B, \hat{B})$ is called auxiliary function of cost function $J(B)$, if it satisfies the following conditions:*

$$Z(B, \hat{B}) \geq J(B), \quad Z(B, B) = J(B) \tag{19}$$

In each iteration t, we optimize over the first parameter:

$$B^{(t+1)} = \arg\min_{B} Z(B, B^{(t)}) \tag{20}$$

By the definition of auxiliary function and minimum, we have $J(B^{(t)}) = Z(B^{(t)}, B^{(t)}) \geq Z(B^{(t+1)}, B^{(t)}) \geq J(B^{(t+1)})$. This method was applied earlier in Expectation Maximization [22] and widely used in NMF literature [16], [19], etc. Due to the lack of space, we have omitted the closed form for our proposed auxiliary function but the following theorems show update rules for B and C variables.

Theorem 1. *The following equations are the multiplicative updates for B variable:*

$$\begin{aligned} &b_{ik} = \hat{b}_{ik}\sqrt{\tfrac{T_{ik}}{T'_{ik}}}, \\ &T_{ik} = 2(K^{-}\hat{B})_{ik} + \textstyle\sum_j x_{ij}\frac{c_{kj}}{\sum_{k'}\hat{b}_{ik'}c_{k'j}},\ (K = \mathcal{Q}(\Gamma^T)) \\ &T'_{ik} = Q_{ik} + 2(H\hat{B}^T)_{ki} + 2(K^{+}\hat{B})_{ik},\ \ (Q = \mathbf{1}_D\mathbf{1}_N^T C^T, H = \mathcal{Q}(\mathbf{1}_r)) \end{aligned} \tag{21}$$

where $\hat{B}$ denotes previous iteration of B variable and notation $[.]^+$ ($[.]^-$) indicates positive (negative) part of a matrix. $\mathbf{1}_D$ denotes a vector of all ones with length D. $\mathcal{Q}(.)$ is a squared function of the argument matrix defined as $\mathcal{Q}(A) = AA^T$. We can introduce the new matrix $\Gamma \in \mathbb{R}^{|U|D\times D}$ in which $|U|$ is neighborhood size and D is number of voxels. Γ is a matrix constituted of the following blocks:

$$\begin{aligned} &\Gamma^T = [\Gamma_1^T, \Gamma_2^T, ..., \Gamma_D^T] \quad \text{where } \Gamma_i \in R^{|U|\times D} \\ &\text{where } [\Gamma_i]_{jl} = \sqrt{\tfrac{w_{il}}{\delta}} \text{ if } k = i \text{ and } [\Gamma_i]_{jl} = -\sqrt{\tfrac{w_{il}}{\delta}} \text{ if } l \in U_i(j) \end{aligned} \tag{22}$$

Proof. Derivation of auxiliary function and multiplicative updates are omitted due to lack of space. For more information, please see our technical support. [1]

Theorem 2. *Following equations are multiplicative updates for C variable:*

$$\begin{aligned} &c_{ik} = \hat{c}_{ik}\sqrt{\tfrac{T_{ik}}{T'_{ik}}}, \\ &T_{ik} = 2(\hat{C}\Lambda_1^{-T})_{ik} + 2(\hat{C}\Lambda_2^{+T})_{ik} + 2\textstyle\sum_l (E^l\hat{C})_{ik} + \sum_j x_{ji}\frac{b_{jk}}{\sum_{k'} b_{jk'}\hat{c}_{k'i}}, \\ &T'_{ik} = 2(\hat{C}\Lambda_1^{+T})_{ki} + 2(\hat{C}\Lambda_2^{-T})_{ki} + M_{ik}, \qquad (M = \mathbf{1}_N\mathbf{1}_D^T B) \end{aligned} \tag{23}$$

Here, N_1 and N_2 are numbers of samples for the first and the second classes respectively and we have assumed that samples from the first class constitute the first N_1 columns of X, and $E^l = e_l e_l^T$ in which e_l is l'th unit vector, $\Lambda_1 = \mathcal{Q}([(I_{N_1} - \frac{1}{N_1}\mathbf{1}\mathbf{1}^T); 0]) + \mathcal{Q}([0; (I_{N_2} - \frac{1}{N_2}\mathbf{1}\mathbf{1}^T)])$ and I_{N_1} is an identity matrix of size N_1 and $\Lambda_2 = \mathcal{Q}([\frac{1}{N_1}I_{N_1}; \frac{-1}{N_2}I_{N_2}])$ and $\mathcal{Q}(.)$ was described earlier.

Proof. Derivation of auxiliary function and multiplicative updates are omitted due to lack of space. For more information, please see our technical support.

3 Results

We tested our approach on MR images of Alzheimer's patients and healthy controls from the ADNI study [2]. The dataset we used for this paper included 60 Normal Control (NC) individuals, 60 individuals with Mild Cognitive Impairment (MCI), and 56 Alzheimer's (AD) disease, whose structural MR scans were analyzed. The data sets included standard T1-weighted MR images acquired sagittally using volumetric 3D MPRAGE with 1.25×1.25 mm in-plane spatial resolution and 1.2 mm thick sagittal slices (8 flip angle). Most of the images were obtained using 1.5 T scanners, while a few were obtained using 3T scanners.

[1] http://www.4shared.com/file/81316860/e2be6088/TechSupport.html
[2] http://www.loni.ucla.edu/ADNI/Data

Images were pre-processed similar to other VBA studies; i.e. AC-PC alignment, skull-removal; and non-rigid registration with a standard coordinate system using a non-rigid registration method [23]. Given deformation field for each individual, a map quantifying the regional distribution of gray matter (GM) was formed for each individual. The map quantifies an expansion (or contraction) to the tissue applied by the transformation to transform the image from the original space to the template space. Consequently, map values in the templates space are directly proportional to the volume of the respective structures in the original brain scan. Although this map can be formed for cerebral fluid (CSF), white matter (WM), and GM, we only used maps corresponding to the GM tissue type.

Ten images were chosen randomly from each group (AD, NC, and MCI) to form the matrix X. Entries of B and C matrices were initialized randomly using uniform random generator on the unit interval. After deriving the basis vectors, columns of B, we can rank them. A ranked basis helps to interpret the result by highlighting the most important features. To get robust results, we applied four different feature ranking methods including: (1) SVM Attribute Selection [15] ;(2) Information Gain Ranking [24] ; (3) Symmetrical Uncertainty [24] ; (4) χ^2 [24]; and then, found consensus (voting) on their results. Fig. 1 shows the top three important basis vectors. Interestingly, the most representative basis for group difference between the AD vs. NC groups is exactly localized at hippocampus which is known to be affected by Alzheimer's disease. Other areas are also very localized in the areas that are either associated with memory or known to be affected by AD.

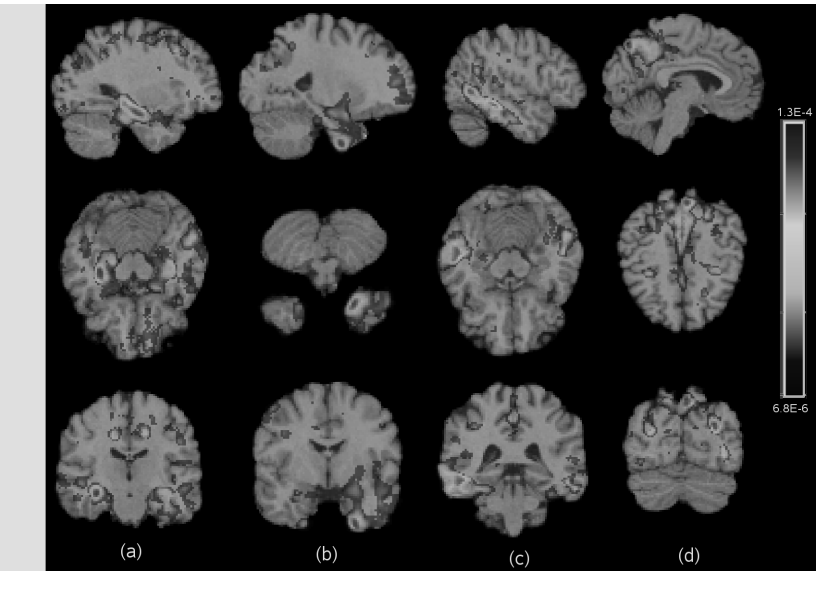

Fig. 1. Three top ranked bases for $r = 50$ and $\lambda = 10^3$: (a) the top ranked basis is localized in hippocampus (b) the second top ranked basis in localized in inferior medial temporal cortex (c) and (d) shows the third top ranked basis being localized in Precuneus and Occipito-parietal association cortex respectively

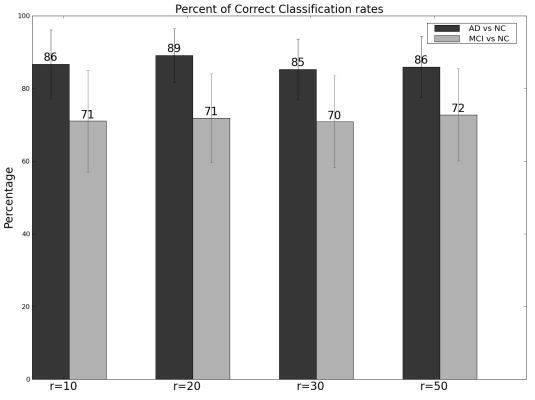

Fig. 2. Comparison of percentages of correct classification rates for the two classification cases when number of bases changes

In order to assess the separability of the new features, we used them for classification. Weka [24] was used to find the best classification strategy in two classification cases: AD vs NC and MCI vs NC. On average, the highest classification rates were obtained when a SVM classifier boosted by the Bagging method is used for AD vs NC and a simple Logistic boosted by Adaboost outperformed other methods for MCI vs NC. Fig. 2 shows the classification rates for the different numbers of basis vectors. It shows the average correct classification rates for ten repetitions of 10-fold cross validations. Classifiers yielded reasonable classification rates for AD vs NC and MCI vs NC cases It is worth mentioning that we used only ten samples from each group (30 samples in total) to build the B matrix. Nevertheless, classification rates are very robust with respect to the changing number of the basis vectors. Besides the fact that there is a narrow difference between definition of AD and MCI cases even for clinicians, we speculate that we can boost this result significantly by using more samples to build the B matrix and using the tissue densities of other tissue types (WM and CSF).

We have also compared the features extracted by our method with features extracted by projecting the data on the principal components (keeping all eigen vectors). The average classification rate with PCA was around %79 but the principal basis vectors were not sparse and hence hard to interpret. In addition, without the Fisher term (15) classification rates droped below %80 although basis vectors were sparse.

We also evaluated the effect of the MRF term introduced in Sect.2.3. As expected, increasing the weight of the MRF term (here we called it λ) leads to a smoother base. Fig.3 depicts the highest ranked basis image for three different values of λ. From left to right, the base becomes smoother and the correct classification rate (for AD vs NC) decreases monotically but not significantly. Eventually in Fig.3(c), the MRF term dominates the other terms and oversmoothes the image however λ was set to a very high value ($\approx 10^4$) to yield such result. This figure shows that our method is robust with regard to choice of weight for the MRF parameter.

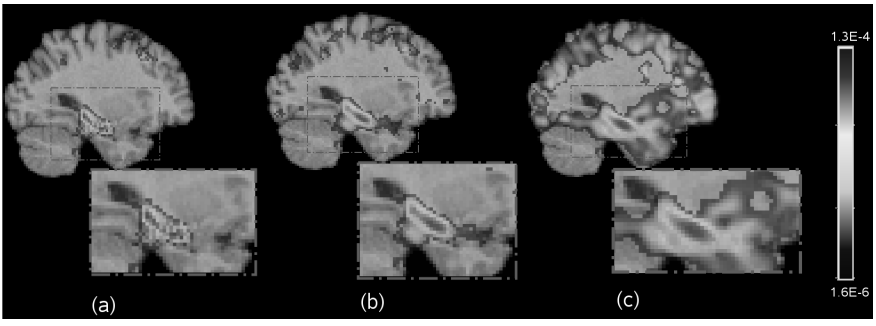

Fig. 3. Effect of the MRF term: (a) for $\lambda = 10$ classification rate was $88.4\% \pm 8.6$, (b) for $\lambda = 10^3$ classification rate was $86.0\% \pm 8.4$, (c) for $\lambda = 10^4$ classification rate was $84.7\% \pm 9.3$

4 Discussion

In this paper, we proposed a novel method based on the NMF framework. Our method is able to produce bases which are simultaneously discriminative and representative of group differences. Moreover, the sparsity of the estimated basis is likely to lead to good generalization of new samples, and to better interpretability of the results via locally-extracted features. The method produces reasonable classification rates between AD patients and normal controls, as well as between normal controls and MCI individuals We plan to improve our results with the current dataset by amending implementation and using tissue density of white matter and CSF. We also plan to apply our method on other datasets and augment it for a vectorial dataset by extending our framework from matrix factorization to tensor factorization.

References

1. Teipel, S.J., Born, C., Ewers, M., Bokde, A.L., Reiser, M.F., Müller, H.J., Hampel, H.: Multivariate deformation-based analysis of brain atrophy to predict alzheimer's disease in mild cognitive impairment. NeuroImage 38(1), 13–24 (2007)
2. Hua, X., et al.: 3D characterization of brain atrophy in alzheimer's disease and mild cognitive impairment using tensor-based morphometry. NeuroImage 41(1), 19–34 (2008)
3. Davatzikos, C., Genc, A., Xu, D., Resnick, S.M.: Voxel-based morphometry using the ravens maps: Methods and validation using simulated longitudinal atrophy. NeuroImage 14(6), 1361–1369 (2001)
4. Wright, I.C., McGuire, P.K., Poline, J.B., Travere, J.M., Murray, R.M., Frith, C.D., Frackowiak, R.S.J., Friston, K.J.: A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. Neuroimage 2(4), 244–252 (1995)
5. Ashburner, J., Friston, K.J.: Voxel-based morphometry-the methods. NeuroImage 11(6), 805–821 (2000)

6. Snook, L., Plewesa, C., Beaulieu, C.: Voxel based versus region of interest analysis in diffusion tensor imaging of neurodevelopment. NeuroImage 34(1), 243–252 (2007)
7. Salmon, E., Collette, F., Degueldre, C., Lemaire, C., Franck, G.: Voxel-based analysis of confounding effects of age and dementia severity on cerebral metabolism in alzheimer's disease. Human Brain Mapping 10(1), 39–48 (2000)
8. Davatzikos, C.: Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. NeuroImage 23, 17–20 (2004)
9. Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C.: Compare: Classification of morphological patterns using adaptive regional elements. IEEE Trans. on Med. Imag. 26(1), 93–105 (2007)
10. Csernansky, J.G., Joshi, S., Wang, L., Haller, J.W., Gado, M., Miller, J.P., Grenander, U., Miller, M.I.: Hippocampal morphometry in schizophrenia by high dimensional brain mapping. Proceedings of the National Academy of Sciences 95(19), 11406–11411 (1998)
11. Thomaz, C., Boardman, J., Counsell, S., Hill, D., Hajnal, J., Edwards, A., Rutherford, M., Gillies, D., Rueckert, D.: A multivariate statistical analysis of the developing human brain in preterm infants. Image and Vision Computing 25(6), 981–994 (2007)
12. Lashkari, D., Vul, E., Kanwisher, N., Golland, P.: Discovering structure in the space of activation profiles in fMRI. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 1015–1024. Springer, Heidelberg (2008)
13. Terriberry, T.B., Joshi, S.C., Gerig, G.: Hypothesis testing with nonlinear shape models. Inf. Process. Med. Imaging 3565(19), 15–26 (2005)
14. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)
15. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422 (2002)
16. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization, 556–562 (2000)
17. Donoho, D., Stodden, V.: When does non-negative matrix factorization give a correct decomposition into parts? In: NIPS, vol. 16, pp. 1141–1148 (2004)
18. Sra, S., Dhillon, I.S.: Technical report, Dept. Computer Science, University of Texas at Austin, Austin, TX 78712, USA (June)
19. Feng, T., Li, S., Shum, H.Y., Zhang, H.: Local non-negative matrix factorization as a visual representation. In: The 2nd International Conference on Development and Learning (2002)
20. Zdunek, R., Cichocki, A.: Blind image separation using nonnegative matrix factorization with gibbs smoothing. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 519–528. Springer, Heidelberg (2008)
21. Wang, Y., Jia, Y., Hu, C., Turk, M.: Fisher non-negative matrix factorization for learning local features. In: Proc. Asian Conf. on Comp. Vision (2004)
22. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, Series B 39(1), 1–38 (1977)
23. Shen, D., Davatzikos, C.: Very high resolution morphometry using mass-preserving deformations and hammer elastic registration. NeuroImage 18, 28–41 (2003)
24. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Elsevier, Amsterdam (2005)