ODVBA: Optimally-Discriminative Voxel-Based Analysis

Tianhao Zhang*, *Member, IEEE,* and Christos Davatzikos, *Senior Member, IEEE,* [†]

***Abstract*—Gaussian smoothing of images prior to applying voxel-based statistics is an important step in Voxel-Based Analysis and Statistical Parametric Mapping (VBA-SPM), and is used to account for registration errors, to Gaussianize the data, and to integrate imaging signals from a region around each voxel. However, it has also become a limitation of VBA-SPM based methods, since it is often chosen empirically and lacks spatial adaptivity to the shape and spatial extent of the region of interest, such as a region of atrophy or functional activity. In this paper, we propose a new framework, named Optimally-Discriminative Voxel-Based Analysis (ODVBA), for determining the optimal spatially adaptive smoothing of images, followed by applying voxel-based group analysis. In ODVBA, Nonnegative Discriminative Projection is applied regionally to get the direction that best discriminates between two groups, e.g., patients and controls; this direction is equivalent to local filtering by an optimal kernel whose coefficients define the optimally discriminative direction. By considering all the neighborhoods that contain a given voxel, we then compose this information to produce the statistic for each voxel. Finally, permutation tests are used to obtain a statistical parametric map of group differences. ODVBA has been evaluated using simulated data in which the ground truth is known and with data from an Alzheimer's disease (AD) study. The experimental results have shown that the proposed ODVBA can precisely describe the shape and location of structural abnormality.**

***Index Terms*—Gaussian smoothing, Statistical Parametric Mapping, Nonnegative Discriminative Projection, Optimally-Discriminative Voxel-Based Analysis, Voxel-Based Morphometry, Alzheimer's disease, ADNI.**

I. Introduction

VOXEL-Based Analysis and Statistical Parametric Mapping (VBA-SPM) [1][18] of imaging data have offered the potential to analyze structural and functional data in great spatial detail, without the need to define a priori regions of interest (ROIs). As a result, numerous studies [7][11][23][33][48][49][50] during the past decade have better investigated brain structure and function in normal and diseased populations, and have enabled the quantification of spatio-temporal imaging patterns.

A fundamentally important aspect of VBA-SPM has been the spatial smoothing of images prior to analysis. Typically, Gaussian blurring of full-width-half-max (FWHM) in the range of 8-16mm is used to account for registration errors, to Gaussianize the data, and to integrate imaging signals from a region, rather than from a single voxel. The effect of this smoothing function is critical: if the kernel is too small for the task, statistical power will be lost and large numbers of false negatives will result in missing many regions that might present group differences in structure and function; if the kernel is too large, statistical power can also be lost by blurring image measurements from regions that display group differences with measurements from regions that have no group difference. In the latter case, spatial localization is also seriously compromised, as significant smoothing blurs the measurements out and often leads to false conclusions about the origin of a functional activation or of brain atrophy. Moreover, a filter that is too large, or that is not matched with the underlying group difference, will also have reduced sensitivity in detecting group differences. As a result, Gaussian smoothing is often chosen empirically, or in an ad hoc fashion, an obvious limitation of such VBA-SPM analyses, in part because of its heuristic nature, and in part because it can lead to overfitting of the data without proper cross-validation or correction for multiple comparisons.

However, the most profound limitation of Gaussian smoothing of images prior to applying the General Linear Model (GLM) [18] is its lack of spatial adaptivity to the shape and spatial extent of a structural or functional region of interest. For example, if atrophy or functional activation in the hippocampus is to be detected, Gaussian smoothing will blur volumetric or activation measurements from the hippocampus with such measurements from surrounding tissues, including the ventricles, the fusiform gyrus, the amygdala and the white matter. Previous work in the literature [13] has shown that spatially adaptive filtering of image data can dramatically improve statistical power to detect group differences. However, little is known about how to optimally define the shape and extent of the smoothing filter, so as to maximize the ability of VBA-SPM to detect group effects.

In this paper, we present a mathematically rigorous framework for determining the optimal spatial smoothing of structural (and potentially functional) images, prior to applying voxel-based group analysis. We consider this problem in the context of determining group differences, and we therefore restrict our experiments to voxel-wise statistical hypothesis testing. However, the mathematical formalism and algorithm are generally applicable to any type of VBA. In order to determine the optimal smoothing kernel, a regional discriminative analysis, restricted by appropriate nonnegativity constraints, is

[†]The work used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The work was supported by the National Institutes of Health under Grant R01AG14971. *Asterisk indicates corresponding author.*

*T. Zhang is with the Section of Biomedical Image Analysis, Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: Tianhao.Zhang@uphs.upenn.edu).

C. Davatzikos is with the Section of Biomedical Image Analysis, Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: Christos.Davatzikos@uphs.upenn.edu).

TABLE I
IMPORTANT NOTATIONS USED IN THE PAPER.

Notation	Description	Notation	Description
X	set of voxels in an image	p	significant level
x	coordinates of a single voxel	ϕ	tuning parameter
k	number of voxels in one neighborhood	τ^2	regularization parameter
$\vec{\theta}$	vector of subvolume	I	identity matrix
Θ	learning set	N_i	number of samples in i^{th} class
$\mathbb{R}^m$	m-dimensional Euclidean space	$\mathbb{N}$	given neighborhood
M	number of voxels in X	μ	balance parameter
$\vec{w}$	nonnegative discriminative direction	γ	controlling parameter
S_W	within-class scatter matrix	B	set of selected neighborhoods
S_B	between-class scatter matrix	F	set of all the neighborhoods
δ_i	discrimination degree	U	set of uncovered voxels
$\vec{m}_i$	class mean	G	graph on voxels
$\vec{e}$	vector with all ones	G_{sub}	submatrix of G

applied to a spatial neighborhood around each voxel, aiming to find the direction (in a space of dimensionality equal to the size of the neighborhood) that best highlights the difference between two groups in that neighborhood. Since each voxel belongs to a large number of such neighborhoods, each centered on one of its neighboring voxels, the group difference at each voxel is determined by a composition of all these optimal smoothing directions. Permutation tests are used to obtain the statistical significance of the resulting ODVBA maps.

This approach is akin to some fundamental principles of signal and image processing, and more specifically to the matched filtering, which states that optimal detection of a signal in the presence of noise is achieved by filtering whose kernel is related to the signal itself. In the context of voxel-based statistical analysis, the "signal" is not known, as it relates to the underlying (unknown) group difference. Therefore, the purpose of our optimization is to actually find the kernel that maximizes the signal detection.

ODVBA has some similarities with the searchlight approach [29][30], however it is significantly different. The searchlight method is basically a local multivariate analysis constrained to the immediate neighborhood of a voxel, whereas ODVBA performs a high-dimensional discriminative analysis using machine learning technique over large neighborhoods, which captures anatomical and functional patterns of larger range, thereby determining the optimally discriminative spatially varying filter. ODVBA also relates to the extensive literature using linear discriminant analysis (LDA) on multivariate patterns of whole brain images. However, implementing the standard LDA directly on the images usually suffers from the singularity problem [14] because the number of images is much smaller than the number of brain voxels. To overcome the problem, Kustra and Strother [31] used the smoothness-constrained, penalized LDA as a tool not only for a strict classification task of positron emission tomography (PEI) images, but also for extracting the activation patterns. Thomaz et al. [45][46] presented a PCA plus Maximum uncertainty LDA that solves the small sample size problem for classification and visual analysis of structural MRIs. Carlson et al. [6] used PCA plus LDA to classify brain activities of different stimulus categories [20][40] and to find which voxels contribute to the activity. Unfortunately, although the singularity problem has been addressed in the above LDA-based methods, a great deal of important information in the images would be lost, since they employ the smoothness constraint or PCA to reduce the dimensionality prior to implementing LDA. However, ODVBA does not have the singularity problem since it never involves the matrix inverse computation and the discriminative analysis is conducted on the data set constructed according to the local neighborhood so it avoids the curse of dimensionality naturally. More importantly, all these three methods attempt to obtain discriminants or a "canonical image", which is a spatial distribution of voxels that maximally differentiates between different experimental conditions, for interpretation of abnormality or activation in the groups. However, the resulting discriminants are then usually analyzed based on visual inspection or simply thresholding without determining a voxel-wise statistical value; therefore, they do not produce the p values for each voxel in the style of traditional SPM, whereas ODVBA does. Finally, ODVBA also employs a non-negativity constraint, which is important as it prohibits canonical images with positive and negative value cancelations, which are often difficult to interpret, especially if a brain region is involved in many canonical images with different weights.

The rest of the paper is organized as follows: Section II describes the general formulation, and its numerical optimization solution. Section III introduces a method of computationally efficient implementation for the ODVBA. Section IV presents a number of experiments with 1) simulated data of known ground truth and 2) structural images of elderly individuals with Alzheimer's disease (AD). These experiments demonstrate that the proposed methodology significantly improves both the statistical power in detecting group differences, and the accuracy with which the spatial extent of the region of interest is determined by VBA-SPM analysis. Section V contains the discussion and conclusion. For convenience, Table I lists important notations used in the paper.

II. THE PROPOSED FRAMEWORK

The proposed framework consists of the following stages:

1) Regional Nonnegative Discriminative Projection. For each voxel, we examine a (typically large) neighborhood centered on it (sometimes referred to as a "subvolume"),

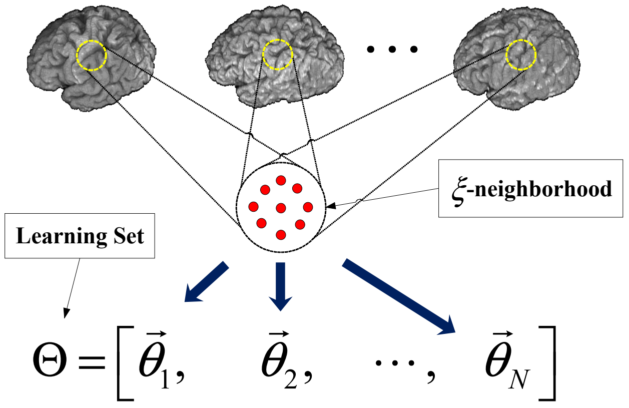
ξ-neighborhood
Learning Set
$\Theta = \left[\vec{\theta}_1, \quad \vec{\theta}_2, \quad \cdots, \quad \vec{\theta}_N \right]$

Fig. 1. Learning set construction.

and use the Nonnegative Discriminative Projection (NDP) to find the direction that best discriminates the two groups being compared. This direction can be viewed as a spatially adaptive filter, which locally amplifies the group differences, since projection is a weighted average of all voxels in the subvolume, i.e. it is a filter.

2) Determining each voxel's statistic. A given voxel belongs to a number of neighborhoods centered on its neighboring voxels, so it may correspond to a collection of values, each being part of a different discriminative direction and reflecting underlying group differences. In order to eventually determine a single value for each voxel that represents the group difference at that spatial location and will be used for statistical analysis, we observe that the contribution of each voxel to a discriminative direction of a neighborhood to which it belongs is given by the respective coefficient of the optimally discriminative direction of that neighborhood. We sum up all these coefficients belonging to a given voxel to determine a quantity that reflects the voxel's discriminating value.

3) Permutation tests. Since assumptions of Gaussianity cannot be made for the derived voxel-wise discriminative measurements, we resort to permutation tests [24][39] to obtain statistical significance (p values).

4) Classification (optional). Although not necessarily part of ODVBA, classification is demonstrated here as an additional use of ODVBA. In particular, along the lines of the COMPARE algorithm [16], the regional clusters showing the highest group differences are used as the input features to be fed to a classifier, e.g., an SVM.

A. Regional Nonnegative Discriminative Projection

For a given voxel x in volume X, we construct its neighborhood $\mathbb{N}$ in which each voxel x_i follows $\|x - x_i\| < \xi$. To render this process computationally efficient when the neighborhood size is large, we randomly select $k-1$ voxels $x_1, \cdots, x_{k-1}$ in this neighborhood and represent this neighborhood using a k dimensional subvolume vector: $\vec{\theta} = [x, x_1, \cdots, x_{k-1}]^T$. Provided that there are N subjects, we can obtain N subvolume vectors which form a data set: $\Theta = [\vec{\theta}_1, \vec{\theta}_2, \cdots, \vec{\theta}_N]$ for learning. The procedure can be illustrated as Fig. 1. If there are M voxels in each subject (i.e. if the template to which all subjects were normalized has M voxels), we will have M learning sets.

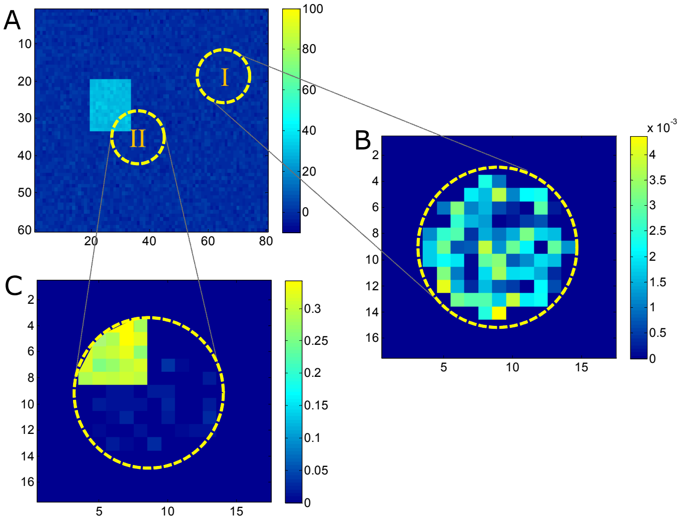

Fig. 2. Illustration of the basic idea of NDP using a toy dataset.

1) The Basic Idea of NDP: The Nonnegative Discriminative Projection (NDP) algorithm which is developed based on Nonnegative Quadratic Programming [42] is used to find the optimal discriminative directions which project the high-dimensional samples onto a 1-dimensional space that maximizes the classification accuracy, and therefore the group differences. NDP is implemented on each learning set. The resultant optimally discriminant vector $\vec{w}$ has a special property: it is nonnegative. This stems from the nonnegativity constraint that is incorporated into the objective function of the NDP method. This constraint is used to help us interpret the group differences. Specifically, our goal is not simply to find an image contrast, prescribed by $\vec{w}$, which distinguishes the two groups, but also requires that this contrast tells us something about the properties of the images we are measuring, e.g. about regional volumetrics or functional activity. We therefore limit ourselves to nonnegative, albeit arbitrarily shaped, local filters, each of which prescribes a regional weighted average of the signal being measured. In a regional volumetric analysis, for example, the optimal regional filter will reflect the shape of the region whose volume is different between two groups. In a functional image analysis, this filter might represent the region whose signal is summed up to reflect the activation. This is in contrast with the traditional methods of feature extraction for pattern classification, which are free to derive any feature obtained by any image filter that maximizes classification accuracy, but are not designed to necessarily measure (interpretable) group differences.

To better illustrate the idea of NDP, we show its results on a toy dataset before we describe the formulation. In this artificial test, we generated images that contained a square (the region of group difference) with intensity that varied from one image to another. We generated two groups of images: the first set of squares had intensities with mean 120.53 and standard deviation 5.79, while the second had 90.36 and 5.72, respectively. Fig. 2A shows the difference of means from the two groups. Fig. 2B shows the $\vec{w}$ obtained from the learning set constructed according to the neighborhood I; it is basically noise with very small values of $\left(\vec{w}\right)_i$, indicating that no local filter can be found that distinguishes the two groups at that neighborhood. Fig. 2C shows the $\vec{w}$ obtained from the learning set constructed according to the neighborhood II; the estimated optimal filter $\vec{w}$ is well aligned with the underlying group difference, within which it has high values. The bottom line here is that, following the formulation to be presented next, the filter that locally amplifies the group difference in this set of images calls for a weighted average of the pixel values within the part of the square that is included in the neighborhood, correctly reflecting the underlying difference. We now present the details of the formulation.

2) The Formulation of NDP: Using one given learning set constructed by the subvolume vectors from all the involved individuals, we probe into neighborhood elements' contributions for discrimination of the two groups. We expect to find such a nonnegative vector $\vec{w}$ to describe the contributions of elements in the neighborhood: the larger the value of $\left(\vec{w}\right)_i$ is, the more the corresponding element $\left(\vec{\theta}\right)_i$ contributes to the discrimination. Equivalently, $\left(\vec{w}\right)_i$ is the i^{th} coefficient of the regional filter denoted by $\vec{w}$. By exploiting $\vec{w}$, the learning set can be projected from the k-dimensional space $\mathbb{R}^k$ onto the 1-dimensional space $\mathbb{R}$ to be optimally classified, such as

$$\Psi = \vec{w}^T\vec{\theta} = \left(\vec{w}\right)_1 x + \left(\vec{w}\right)_2 x_1 + \cdots + \left(\vec{w}\right)_k x_{k-1}. \tag{1}$$

where, Ψ is the 1-dimensional projection of $\vec{\theta} \in \mathbb{R}^k$. We expect that the two classes will be separated as much as possible along $\vec{w}$, and at the same time the samples from the same class get more compact.

A measure of the separation between the two classes is the distance of projected class means:

$$\begin{aligned}\left(\tilde{m}_1 - \tilde{m}_2\right)^2 &= \left(\vec{w}^T\vec{m}_1 - \vec{w}^T\vec{m}_2\right)^2 \\ &= \vec{w}^T\left(\vec{m}_1 - \vec{m}_2\right)\left(\vec{m}_1 - \vec{m}_2\right)^T\vec{w} \\ &= \vec{w}^T S_B \vec{w}\end{aligned} \tag{2}$$

where, $\tilde{m}_i = \frac{1}{N_i}\sum_{\Psi \in C_i}\Psi$; C_i means the i^{th} class, $i = 1, 2$; N_i denotes the number of samples in C_i; $\vec{m}_i = \frac{1}{N_i}\sum_{\vec{\theta} \in C_i}\vec{\theta}$; $S_B = \left(\vec{m}_1 - \vec{m}_2\right)\left(\vec{m}_1 - \vec{m}_2\right)^T$.

And, the intra-class compactness is defined as follows:

$$\begin{aligned}&\sum_{i=1}^{2}\sum_{\Psi \in C_i}\left(\Psi - \tilde{m}_i\right)^2 = \sum_{i=1}^{2}\sum_{\vec{\theta} \in C_i}\left(\vec{w}^T\vec{\theta} - \vec{w}^T\vec{m}_i\right)^2 \\ &= \vec{w}^T\left(\sum_{i=1}^{2}\sum_{\vec{\theta} \in C_i}\left(\vec{\theta} - \vec{m}_i\right)\left(\vec{\theta} - \vec{m}_i\right)^T\right)\vec{w} \\ &= \vec{w}^T S_W \vec{w}\end{aligned} \tag{3}$$

where, $S_W = \sum_{i=1}^{2}\sum_{\vec{\theta} \in C_i}\left(\vec{\theta} - \vec{m}_i\right)\left(\vec{\theta} - \vec{m}_i\right)^T$.

S_B and S_W are called the between-class scatter matrix and the within-class scatter matrix separately, according to the classic Fisher Linear Discriminant Anaylsis [14] in which the criterion function is based on the generalized Rayleigh quotient. Herein we consider S_B and S_W under the formulation of quadratic programming which is amenable to the nonnegative constraint as follows:

$$\begin{aligned}&J(\vec{w}) = \min_{\vec{w}} \vec{w}^T A\vec{w} - \mu\vec{e}^T\vec{w} \\ &\text{subject to } (\vec{w})_i \geq 0,\ i = 1, \cdots, k,\end{aligned} \tag{4}$$

where, $A = (\gamma S_W - S_B + (|\lambda_{min}| + \tau^2)I)$; γ is the controlling parameter; $|\lambda_{min}|$ is the absolute value of the smallest eigenvalue of $\gamma S_W - S_B$; $\tau^2 << 1$ acts as the regularization parameter; I is the identity matrix; $\vec{e}$ is the vector with all ones; the second term $\vec{e}^T\vec{w}$ is used to achieve $\sum_{i=1}^{k}(\vec{w})_i > 0$ which means the solutions of $(\vec{w})_i$ are not all zeros under the nonnegative constraint; μ is the balance parameter.

Theorem 1. A is a positive definite matrix.

The proof is in the Appendix A.

Since A is positive definite, $J(\vec{w})$ is a convex function and has the unique global minimum. We solve the above optimization problem using the Nonnegative Quadratic Programming (NQP) [42]. According to [42], let A^+ and A^- denote the nonnegative matrices described as:

$$A^+ = \begin{cases} A_{ij}, & \text{if } A_{ij} > 0, \\ 0, & \text{otherwise.}\end{cases} \tag{5}$$

and

$$A^- = \begin{cases} |A_{ij}|, & \text{if } A_{ij} < 0, \\ 0, & \text{otherwise.}\end{cases} \tag{6}$$

It is clear that $A = A^+ - A^-$. According to the two nonnegative matrices, the objective function in (4) can be written as the combination of three terms:

$$J(\vec{w}) = J_a(\vec{w}) + J_b(\vec{w}) - J_c(\vec{w}). \tag{7}$$

where

$$\begin{aligned}J_a(\vec{w}) &= \vec{w}^T A^+\vec{w}, \\ J_b(\vec{w}) &= -\mu\vec{e}^T\vec{w}, \\ J_c(\vec{w}) &= \vec{w}^T A^-\vec{w}.\end{aligned} \tag{8}$$

We define their derivatives as follows:

$$\begin{aligned}\frac{\partial J_a}{\partial\vec{w}} &= 2A^+\vec{w}, \\ \frac{\partial J_b}{\partial\vec{w}} &= -\mu\vec{e}, \\ \frac{\partial J_c}{\partial\vec{w}} &= 2A^-\vec{w}.\end{aligned} \tag{9}$$

Note that the partial derivatives on J_a and J_c above are guaranteed to be nonnegative because of the nonnegativity of $\vec{w}$.

Using the above derivatives, multiplicative updates rule which does not involve the learning rates is introduced to minimize the objective function iteratively:

$$(\vec{w})_i \leftarrow \left(\frac{(\mu\vec{e})_i + \sqrt{(\mu\vec{e})_i^2 + 16(A^+\vec{w})_i(A^-\vec{w})_i}}{4(A^+\vec{w})_i}\right)(\vec{w})_i, \tag{10}$$

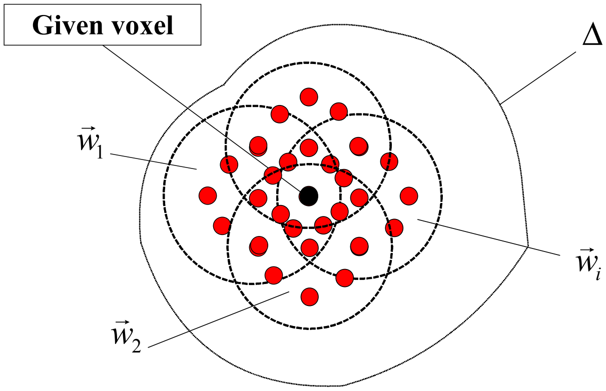

Fig. 3. A given voxel belonging to a set of neighborhoods.

where $i = 1, \cdots, k$. Equation (10) means that all the elements in $\vec{w}$ are updated in parallel. Since $A^{+}\vec{w} \geq 0$ and $A^{-}\vec{w} \geq 0$, the updated $\vec{w}$ in(10) is always nonnegative.

Theorem 2. The function of $J(\vec{w})$ in Equation (4) decreases monotonically to the value of its global minimum under the multiplicative updates in Equation (10).

The proof is in Appendix B.

B. *Determining each voxel's statistic*

For all the M voxels in one volume, we have M discriminative directions, each applied to a different neighborhood, as described in Section II-A. For a given voxel x, we obtain a list of $(\vec{w})_i$ values from different $\vec{w}$ since x may belong to a number of neighborhoods, as shown in Fig. 3. (Recall that for each neighborhood to which x belongs, the respective coefficient in $\vec{w}$ reflects the discrimination power of this voxel in terms of the pattern seen in that neighborhood.). To quantify the group difference measured at voxel x, we use the following function, termed $discrimination\ degree$, which relates to the effect size [8]:

$$\delta = \left(\frac{|\tilde{m}_1 - \tilde{m}_2|}{\sqrt{\sum\limits_{i=1}^{2} \sum\limits_{\Psi \in C_i} (\Psi - \tilde{m}_i)^2}} \sqrt{N_1 + N_2 - 2} \right)^{\phi}, \qquad (11)$$

where, ϕ is a tuning parameter aiming to reduce potential outliers in the dataset. Let $\Delta = \{\mathbb{N} | x \in \mathbb{N}\}$ denote the set of neighborhoods that the given voxel x belongs to, then we define the group difference on x by summing up contributions from all neighborhoods to which it participates:

$$S_x = \sum_{\mathbb{N} \in \Delta} \delta_{\mathbb{N}} \left(\vec{w}_{\mathbb{N}}\right)_i, \ i \in \{1, \cdots, k\}, \qquad (12)$$

where, $\vec{w}_{\mathbb{N}}$ denotes the coefficients corresponding to voxels in $\mathbb{N}$, $(\vec{w}_{\mathbb{N}})_i$ denotes that x is the i^{th} element in $\mathbb{N}$, and $\delta_{\mathbb{N}}$ which acts as the weight for $\vec{w}_{\mathbb{N}}$ denotes the $discrimination\ degree$ achieved in neighborhood $\mathbb{N}$ and is defined in (11). S_x will serve as the statistic reflecting group differences on the respective voxel x, and will be used next to determine statistical significance. Higher values of S_x reflect stronger group differences.

It is worth noting that, in ODVBA, one learning set is constructed according to one given neighborhood (subvolume vector) by different individuals. Actually, based on one learning set, the obtained coefficient $\vec{w}$ has only one direction (group1>group2 / group2>group1) to maximize the classification accuracy. When ODVBA handle data that have both possible positive and negative differences, it does not need to calculate the $\vec{w}$ twice using Nonnegative Discriminative Projection described in Section II-A, but it has to determine two different statistical values S_x for both positive and negative cases, before implementing permutation test respectively. In particular, only having the coefficient $\vec{w}$, by which the projected class means $\tilde{m}_{group1} > \tilde{m}_{group2}$ (referring to (2)), involved in (12), we can get the statistical value for positive case S_x^P. Conversely, only use $\vec{w}$ by which the projected class means $\tilde{m}_{group1} < \tilde{m}_{group2}$, we can get the S_x^N for negative case.

Moreover, for one given neighborhood, the optimal coefficients can only reflect the positive (assuming group1>group2) signals if the projected class means $\tilde{m}_{group1} > \tilde{m}_{group2}$ in this subvolume. However, the negative signals will never be ignored if they are really significant, because they may be highlighted in other random neighborhood where they are stronger than the positive ones, so that $\tilde{m}_{group1} < \tilde{m}_{group2}$.

C. *Permutation tests*

Assuming that the null hypothesis is that there is no difference between the two groups, the statistical significance can be assessed by comparison with the distribution of values obtained when the labels are randomly permuted [24][39]. In particular, we randomly assign the subjects into two groups, and then implement Section II-A and Section II-B to calculate the statistic for each voxel. The above relabeling is repeated N_p times. For one given voxel, let S_0 denote the statistic value obtained under the initial class labels, and S_i, $i = 1, \cdots, N_p$ denote the ones obtained by relabeling. The p value for the given voxel is calculated according to:

$$p = \sum_{i=1}^{N_p} \left[u(S_i - S_0)\right] / N_p, \qquad (13)$$

where,

$$u(t) = \begin{cases} 1, & \text{if } t \geq 0, \\ 0, & \text{otherwise.} \end{cases} \qquad (14)$$

III. Computationally Efficient Implementation

The original method described in Section II is based on voxel-wise computation, an extreme solution that uses all the neighborhoods corresponding to all the M voxels. Unfortunately, this approach would be complicated and computationally expensive. The simplest way to reduce the computational complexity is to randomly select a subset of the neighborhoods to represent all the neighborhoods, and then the number of the involved learning sets can be reduced. However, with this method, the selected neighborhoods may not cover all the

TABLE II
ALGORITHM OF GREEDY NEIGHBORHOODS COVER

```
Input: $X, F, a, k$
Output: $B$
1. Select arbitrary $\mathbb{N}_i \in F$
2. $B \leftarrow \{\mathbb{N}_i\}$
3. $U \leftarrow X - \mathbb{N}_i$
4. while $U \neq \varnothing$
5.    do selet $\mathbb{N}_j \in F$ that maximizes $|\mathbb{N}_j \bigcap U|$
           under the condition $|\mathbb{N}_j \bigcap U| \leq (1-a) \cdot k$
6.       $U \leftarrow X - \mathbb{N}_j$
7.       $B \leftarrow B \bigcup \{\mathbb{N}_j\}$
8. end while
```

TABLE III
ALGORITHM OF GRAPH-ASSISTANTED GREEDY NEIGHBORHOODS COVER

```
Input: $X, F, a, k, G$
Output: $B$
1. Select arbitrary $\mathbb{N}_i \in F$
2. $B \leftarrow \{\mathbb{N}_i\}$
3. $U \leftarrow X - \mathbb{N}_i$
4. while $U \neq \varnothing$
5.    do $\tilde{B} = \bigcup_{\mathbb{N} \in B} \mathbb{N}$
6.       get $L$ according to (19)
7.       $G_{sub} = G\left(L, \tilde{B}\right)$
8.       $T = G_{sub} \cdot \vec{e}$
9.       let $i_{min}$ denote the index of element which is the
         minimum in $T$ wich constraint: $T(i_{min}) \geq a \cdot k$
10.      $U \leftarrow X - \mathbb{N}_{L\{i_{min}\}}$
11.      $B \leftarrow B \bigcup \{\mathbb{N}_{L\{i_{min}\}}\}$
12. end while
```

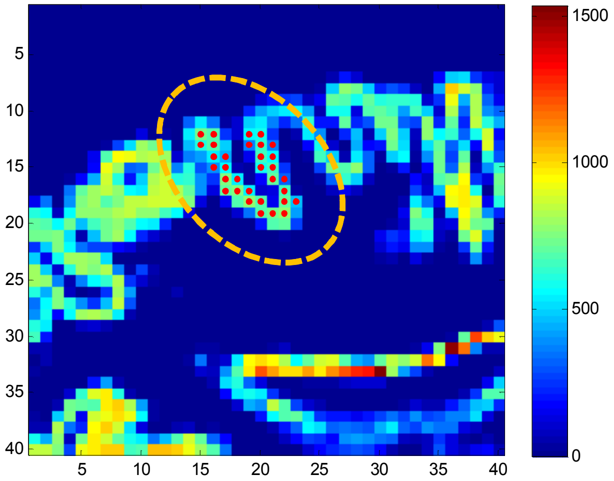

Fig. 4. Location of simulated atrophy (red points in the dashed ellipse) in the "Blurring" case. Different colors in the background reflect different RAVENS values, that is, different tissue densities.

voxels, and therefore, it is not guaranteed that statistics can be derived for each voxel. In this section, we use the Greedy Algorithm [10] to select the subset of the neighborhoods which include all the voxels and at the same time overlap each other under a certain rate. Moreover, to facilitate the greedy search, we introduce a new approach named Graph-assisted Greedy Neighborhoods Cover, which uses a graph to assist in seeking one neighborhood in each iteration of the greedy algorithm.

A. Minimum Subset of the Neighborhoods for ODVBA

We select the minimum subset of the neighborhoods B as the replacement of all the ones F, where $B \subseteq F$, and then the learning sets are constructed corresponding to the selected neighborhoods respectively. To get a reasonable statistic value for each voxel after obtaining the discriminative directions using NDP, the selected neighborhoods are expected: i) to cover all the voxels in X, that is,

$$X = \bigcup_{\mathbb{N} \in B} \mathbb{N}; \tag{15}$$

ii) to overlap each other under at least an overlapping factor [51] a $(0 < a < 1)$, that is,

$$\left| \mathbb{N} \bigcap \left(\bigcup_{Z \in B \backslash \{\mathbb{N}\}} Z \right) \right| \geq a \cdot |\mathbb{N}| = a \cdot k \tag{16}$$

for each $\mathbb{N}$ in B.

B. The Traditional Greedy Neighborhoods Cover Algorithm

The problem introduced in Section III-A can be solved using the greedy algorithm [10][51], which works by picking $\mathbb{N}$ that covers the maximum number of remaining elements that are uncovered under certain constraint in each iteration. Let U contain all the remaining uncovered elements at each iteration. The algorithm chooses the neighborhood $\mathbb{N}_i$ that overlaps with the selected neighborhoods in B and covers as many uncovered elements in U as possible. Then, the selected $\mathbb{N}_i$ is put into B and the elements covered by $\mathbb{N}_i$ is removed from U. It finally terminates when U is empty. The algorithm is summarized in Table II.

C. Graph-assisted Greedy Neighborhoods Cover algorithm

It is worth noting that Line 5 of the algorithm in Table II may be time-consuming because it picks $\mathbb{N}_i$ from all the M neighborhoods without using the neighbor information for searching in each iteration. We propose a Graph-assisted Greedy Algorithm which select $\mathbb{N}_i$ by implementing simple operations on the small sub-block of the graph in each iteration. The algorithm is described as follows.

A graph G is introduced to model the neighbor relationship between every two voxels in X. Each sample point in X is a vertex of the graph. An edge is put from x_i to x_j if x_j is located in the neighborhood centered at x_i . That is:

$$G(i,j) = \begin{cases} 1, & \text{if } x_j \in \mathbb{N}_i, \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

Obviously, each row of G represents one neighborhood. Let $\tilde{B}$ denote the set of the elements covered by the neighborhoods from B, that is $\tilde{B} = \bigcup_{\mathbb{N} \in B} \mathbb{N}$, use I_B to denote the index set of the selected neighborhoods, and we can get the sub-matrix from G as follows:

$$G_{sub} = G\left(L, \tilde{B}\right), \tag{18}$$

where set L follows

$$1)\ L \subset X;$$

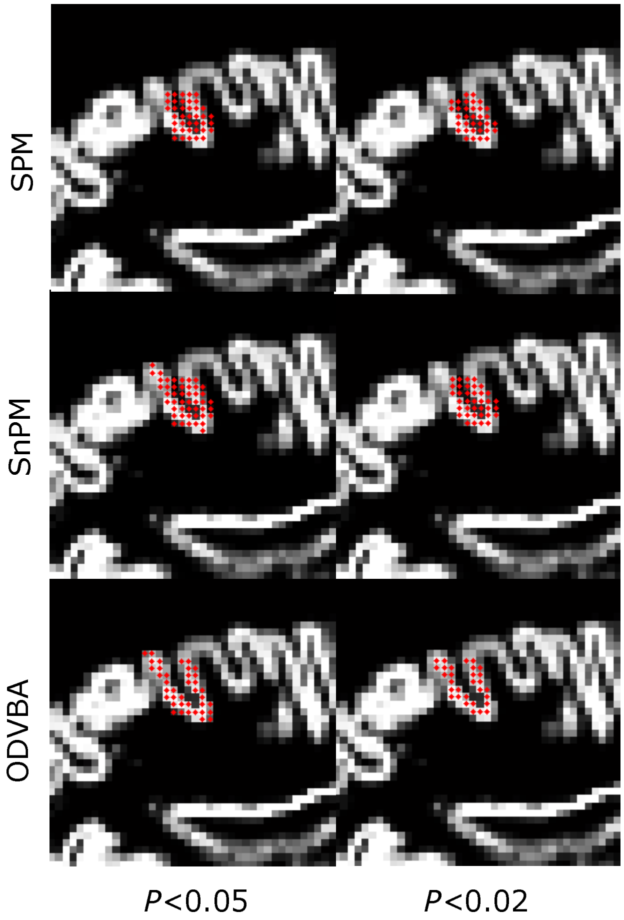

Fig. 5. The maps of p values from the "Blurring" case. The background is the RAVENS map of a sample, appearing as a gray image, and the overlapping areas are the red regions obtained by different p value thresholds.

$$2)\ L \cap I_B = \varnothing;$$

$$3)\ G_{sub}\left(i, \tilde{B}\right) \cdot \vec{e} \neq k, i \in L;$$

$$4)\ G_{sub}\left(i, \tilde{B}\right) \neq 0, i \in L. \quad (19)$$

Among the terms in (19), 1) means that the rows of G_{sub} come from G; 2) means that the neighborhoods to which the rows of G_{sub} correspond have never been selected to B; 3) means the sum of each row in G_{sub} are not equal to k, that is, the neighborhood which are fully covered by the $\tilde{B}$ are removed; 4) means the neighborhoods having no intersection with $\tilde{B}$ are removed.

Compute

$$T = G_{sub} \cdot \vec{e}, \quad (20)$$

and we can know $T(i)$ is the number of elements covered by $\tilde{B}$, in the $L\{i\}^{th}$ neighborhood. Let i_{min} denote the index of element which is the minimum in T with constraint: $T(i_{min}) \geq a \cdot |\mathbb{N}| = a \cdot k$. Then the selected $\mathbb{N}$ in this iteration should be $\mathbb{N}_{L\{i_{min}\}}$. The proposed algorithm is summarized in Table III.

IV. Experiments and Results

In this section, we designed two different kinds of experiments to evaluate the performance of ODVBA compared with the original SPM [1][44] and nonparametric permutation based SPM (referred to SnPM) [24][39]. In the first experiment, we applied the three methods to the simulated atrophy data in which we know the ground truth. Here, it is easy for us to evaluate the accuracy and statistical power of ODVBA, in comparison with SPM and SnPM. The second experiment demonstrated the success of our method on the real data of AD patients from ADNI(www.loni.ucla.edu/ADNI).

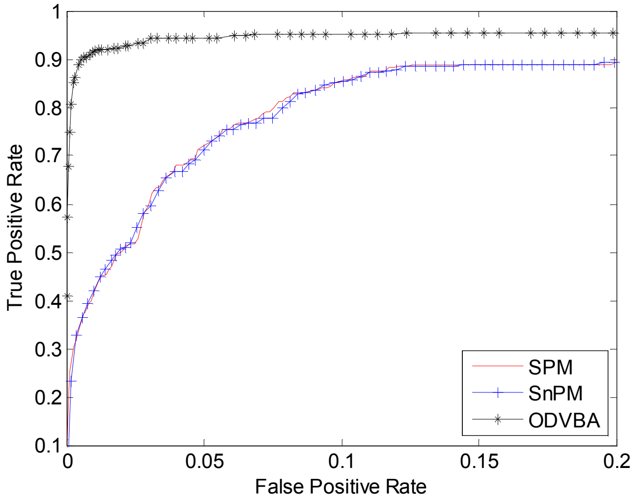

Fig. 6. ROC curve of the "Bluring" case.

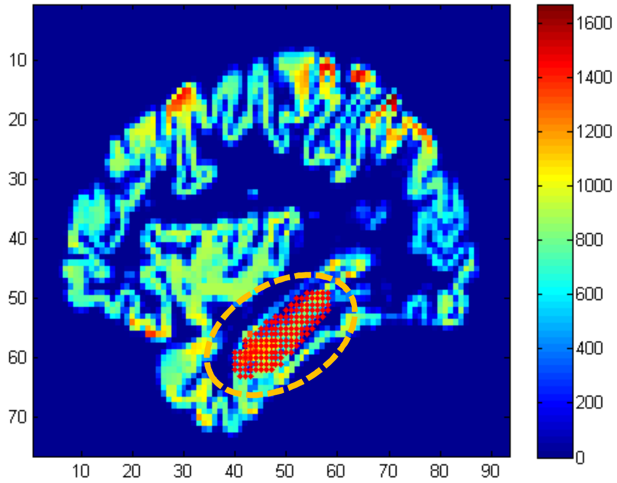

Fig. 7. Location of simulated atrophy (red points in the black dashed ellipse) in the "Different degrees" case. Different colors in the background reflect different RAVENS values, that is, different tissue densities.

A. Simulated Atrophy Data

1) "Blurring" case: The dataset consisted of real MRI scans of 60 normal controls, with a relatively small age range, obtained from ADNI. The description of the MR Image acquisition and pre-processing protocol can be found in Section IV-B. Finally, RAVENS maps [13] which quantify the regional distribution of GM, WM, and CSF, are formed for each tissue type. In this test, RAVENS of GM is employed. Next, 30 samples were randomly picked and manipulated to introduce

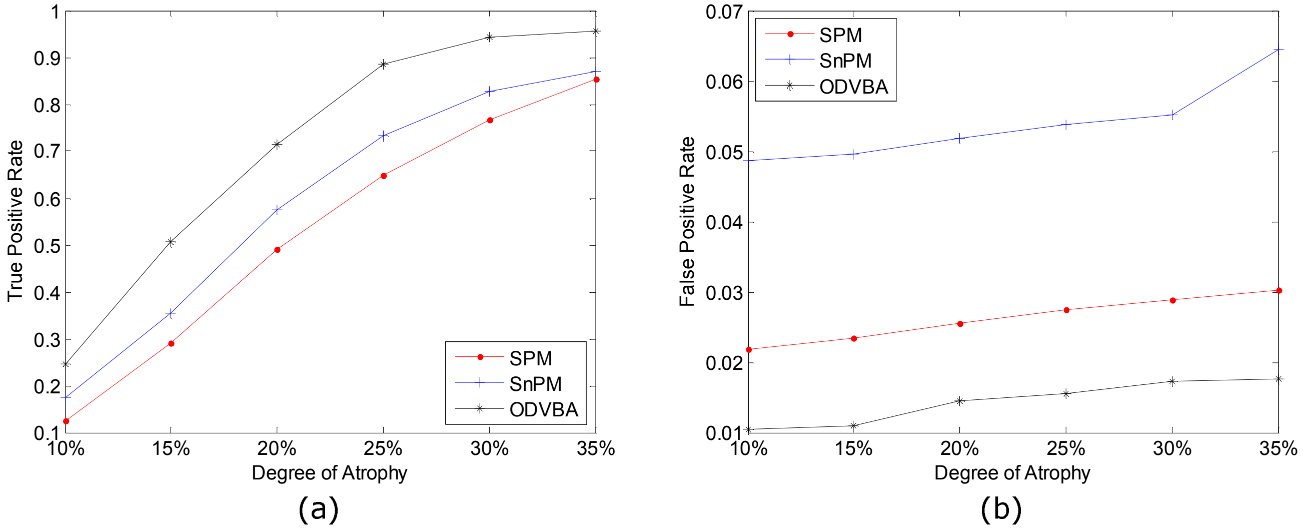

Fig. 8. Performance vs. the degree of atrophy in the “Different degrees” case . (a) TPR vs. the degree of atrophy; (b) FPR vs. the degree of atrophy.

TABLE IV
TPR AND FPR VALUES WITH THE SPECIFIC p VALUE THRESHOLDS IN THE “BLURRING” CASE.

		$P < 0.05$	$P < 0.02$
SPM	TPR	0.4167	0.3194
	FPR	0.0094	0.0028
SnPM	TPR	0.5069	0.3924
	FPR	0.0196	0.0076
ODVBA	TPR	0.8646	0.7535
	FPR	0.0032	0.001

30% atrophy with “U”-like shape in the specific region (a gyrus). In other words, the corresponding RAVENS values in that region are decreased by 30%. Fig. 4. Shows one selected slice of RAVENS map and the location of the atrophy.

Regarding the simulated data as a group of patients and the remaining data as the control group, we conduct SPM, SnPM and ODVBA for the group analysis. The smoothing size (FWHM of the involved Gaussian filter) in SPM and SnPM is 7mm. The parameters in ODVBA are set as $\xi = 15, \phi = 1, \mu = 1, \gamma = 10^{-5}, \tau^2 = 10^{-5}$. The experimental results are shown in Fig. 5. The background is the RAVENS map of a sample, appearing as a gray image, and the overlapping areas are the red regions obtained by different p value thresholds (p <0.05, 0.02). We can see that the proposed ODVBA method precisely describes the “U”-like shape of the atrophy, while SPM and SnPM do not. In SPM and SnPM, the images are blurred so that the results are not accurate.

For the simulated data, since we know the ground truth, we employ several types of metrics to evaluate ODVBA, compared with SPM and SnPM. We denote the significant voxels obtained from the three different methods as V_R, the ground truth voxels as V_G, and all the voxels involved in analysis as V_A. True Positive Rate (TPR) and False Positive Rate (FPR), two commonly used assessment metrics, are defined as follows:

$$TPR = \frac{\#\left(V_R \bigcap V_G\right)}{\#\left(V_G\right)};$$
$$FPR = \frac{\#\left(V_R \bigcap \left(V_A \backslash V_G\right)\right)}{\#\left(V_A \backslash V_G\right)}, \tag{21}$$

where $\#(\cdot)$ means the number of voxels.

We varied the significance level of the group analysis from [0, 1] providing a series of TPR/FPR to build the receiver-operating characteristics (ROC) curve [35], as demonstrated in Fig. 6. We also list the TPR and FPR values with different specific p value thresholds (p <0.05, 0.02) in Table IV. We can see that, for the “Blurring” case, SPM and SnPM not only produce worse accuracies but also suffer from high false positive errors. Otherwise, ODVBA offers a globally better result since for any FPR, the corresponding TPR is higher.

2) “Different degrees” case: The data used in this experiment contain 60 normal controls which also come from ADNI. The purpose is to evaluate the capabilities of three methods for detecting the simulated atrophies with different degrees. As well as in the “Blurring” case, RAVENS maps are finally created to represent the tissue density (which is proportional to regional volume, at that location). We randomly selected 30 volumes of GM and introduced the atrophy of 10%, 15%, 20%, 25%, 30%, and 35% (RAVENS values are reduced by 10%-35%) in the region around Hippocampus, separately. Fig. 7 shows one selected slice and the location of the simulated atrophy. The remaining 30 samples were regarded as control. The smoothing size in SPM and SnPM is 8mm. The parameters in ODVBA are set as $\xi = 15, \phi = 1, \mu = 1, \gamma = 10^{-5}, \tau^2 = 10^{-5}$. Fig. 8 demonstrates the performances (TPR and FPR) of the three involved methods versus the different degrees of atrophy. The p value which is used to get the significant region is 0.05. We can see that ODVBA has higher TPR but lower FPR on each dataset with different degrees of atrophy than SPM and SnPM. In Fig. 9, we provide some representative slices on which we plot the ground truth and the detected significant regions obtained by the three methods. It

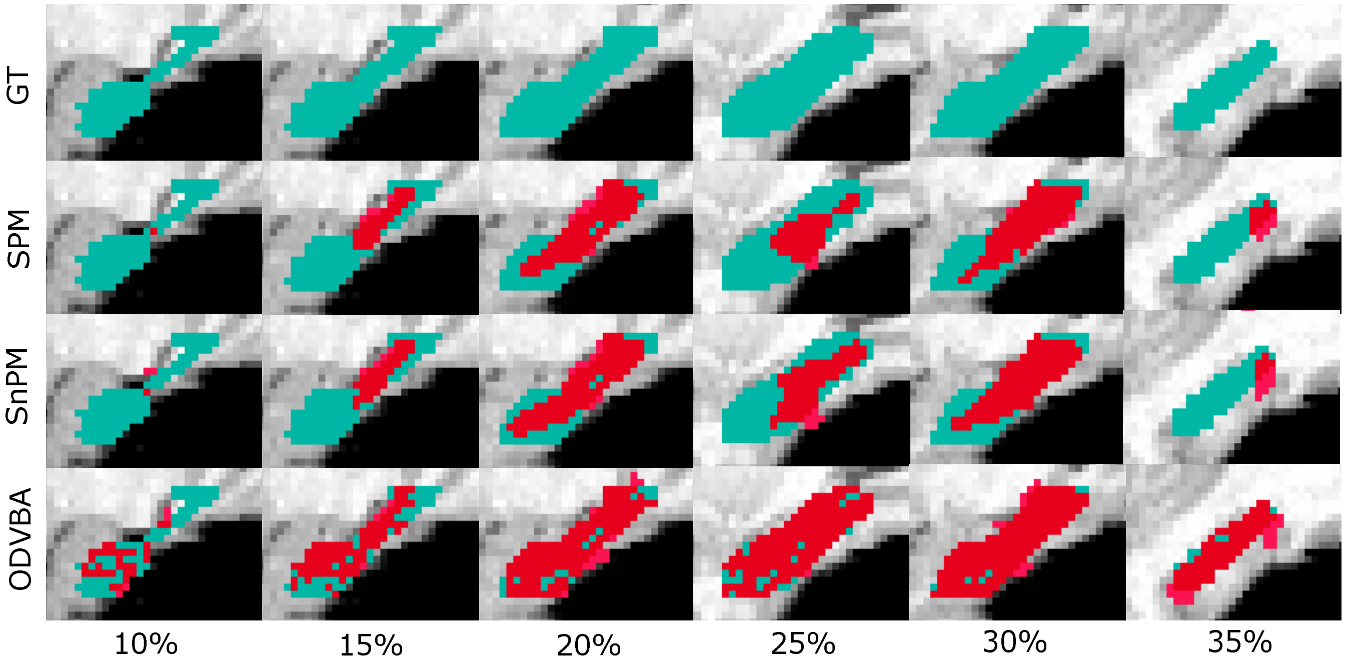

Fig. 9. Respective slices with ground truth and significant regions. The background is the MR image of the template, appearing as a gray image, the overlapping green color areas are the ground truth, and the overlapping red color areas are the detected significant region. The results in 1^{st}-6^{th} columns are obtained from data with 10%-35% atrophy, respectively. The results in 1^{st}-4^{th} columns are threshold by p<0.05; The results in 5^{th} and 6^{th} columns are are threshold p <0.005.

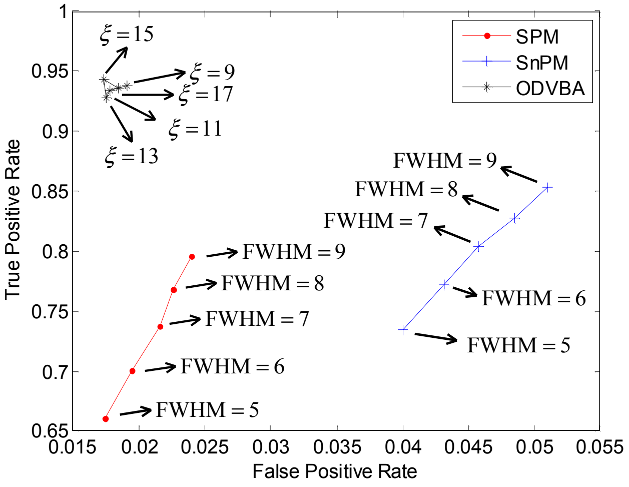

Fig. 10. Performances of three methods with different kernel sizes.

is shown that ODVBA is much more powerful to detect the true atrophy hidden in the data.

3) Effect of kernel size: In this section, we study the effect of the kernel size on the performance of the three methods, using the simulated data with 30% atrophy in the "Different degrees" case. For SPM and SnPM, the kernel size means the FWHM of the involved Gaussian filter, which is changed from 5mm to 9mm. For ODVBA, the kernel size means the size of neighborhood with varying ξ from 9mm to 17mm, with an interval of 2mm. By changing the kernel size, we can get a series of different results for the three different methods. Since the ground truth is known, we plot the corresponding ROC curve in Fig. 10. As shown, the performance of ODVBA is stable. However, the sensitivity/specificity of SPM and SnPM vary a lot with different kernel sizes and are companied with high FPR increasing.

4) Effect of discrimination degree: In this section, we look into the effect of the $discrimination\ degree$ on the performance of ODVBA, using the simulated data with 30% atrophy in the "Different degrees" case. By varying the tuning parameter ϕ described in (11), we can get a series of different resulting significant maps of ODVBA. Note that larger ϕ means the coefficient $(\vec{w})_i$ will be combined with a stronger $discrimination\ degree$ to determine the statistical value (referring to (12)), and $\phi = 0$ means that only $(\vec{w})_i$ is used. Since the ground truth is known, we plot the corresponding TPR and FPR versus ϕ in Fig. 11. As shown, along with the increasing of ϕ, the power of detecting the true positives is gradually enhanced, accompanied by low level FPR.

B. Real AD data

With the development of computer-aided diagnosis, Alzheimer's disease has attracted a lot of attention [3][17][27][28][37][47] in the community of neuroimage. The data used in this experiment was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI), which has recruited approximately 800 adults, ages 55 to 90, including 200 with Normal Control (NC), 400 with mild cognitive impairment (MCI) and 200 with AD. For up-to-date information, see www.adni-info.org.

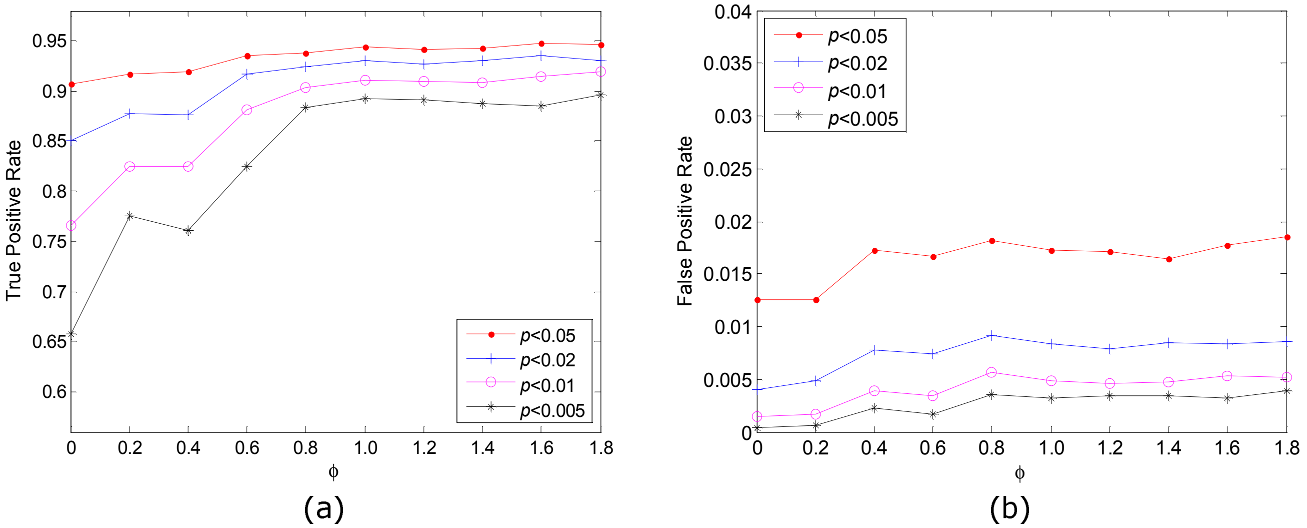

Fig. 11. Performance of ODVBA vs. ϕ. (a) TPR vs. ϕ ; (b) FPR vs. ϕ.

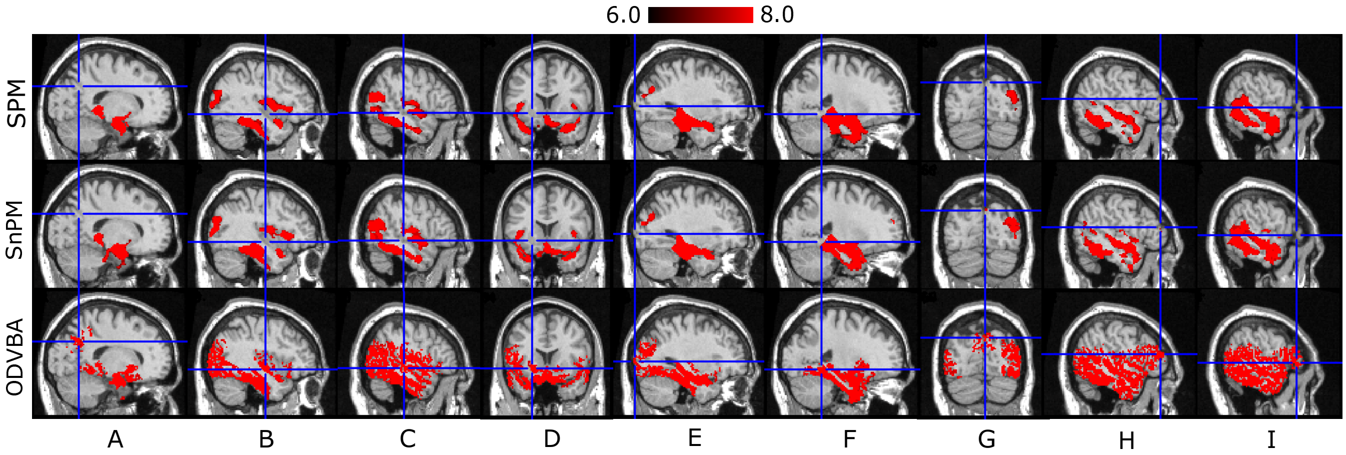

Fig. 12. Representative sections with regions of relatively reduced GM in AD compared to NC. The scale indicates the $-log(p)$ values.

The goal of our study is to conduct complete evaluations of the structural MR images, and identify potentially complex spatial patterns of brain atrophy in AD patients, compared with NC subjects. We randomly selected 100 subjects from the ADNI cohort. This included 50 NCs and 50 ADs, whose MRI scans were then analyzed. The datasets included standard T1-weighted images with varying resolutions. Adhering to volumetric 3D MPRAGE or equivalent protocols, only images obtained using 1.5 T scanners were used. The sagittal images were preprocessed according to a number of steps detailed on the ADNI website, which corrected for field inhomogeneities and image distortion, and were resliced to axial orientation. Images were preprocessed according to the following steps. 1) Alignment of the brain with the ACPC plane; 2) Removal of extra-cranial material (skull-stripping); 3) Tissue segmentation into grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF), using [41]; 4) High-dimensional image warping [43] to a standardized coordinate system, a brain atlas (template) that was aligned with the MNI coordinate space [26]; 5) Formation of tissue density maps, i.e. RAVENS maps, typically used in the modulated SPM analysis [12] and in [2].

RAVENS maps quantify the regional distribution of GM, WM, and CSF, since one RAVENS map is formed for each tissue type. RAVENS values in the template's (stereotaxic) space are directly proportional to the volume of the respective structures in the original brain scan [12]. Therefore, regional volumetric measurements and comparisons can be performed via measurements and comparisons of the RAVENS maps. In this experiment, we used GM for evaluation purposes.

1) Resulting Significant Maps: Based on the measurements of tissue density maps, we compared the performance of SPM, SnPM, and our proposed ODVBA method. For SPM and SnPM, smoothing is performed using 8 mm FWHM kernel. For ODVBA, the parameters are set as follows: $\xi = 15, \phi = 1, \mu = 1, \gamma = 10^{-5}, \tau^2 = 10^{-5}$. Both SnPM and ODVBA were implemented with 2000 permutations. Fig. 12 shows some selected sections from the results (with the p value $<$ 0.001 threshold) of SPM, SnPM, and our ODVBA, respectively. We can see that the results of ODVBA not only reflect significant GM loss in AD compared with NC,

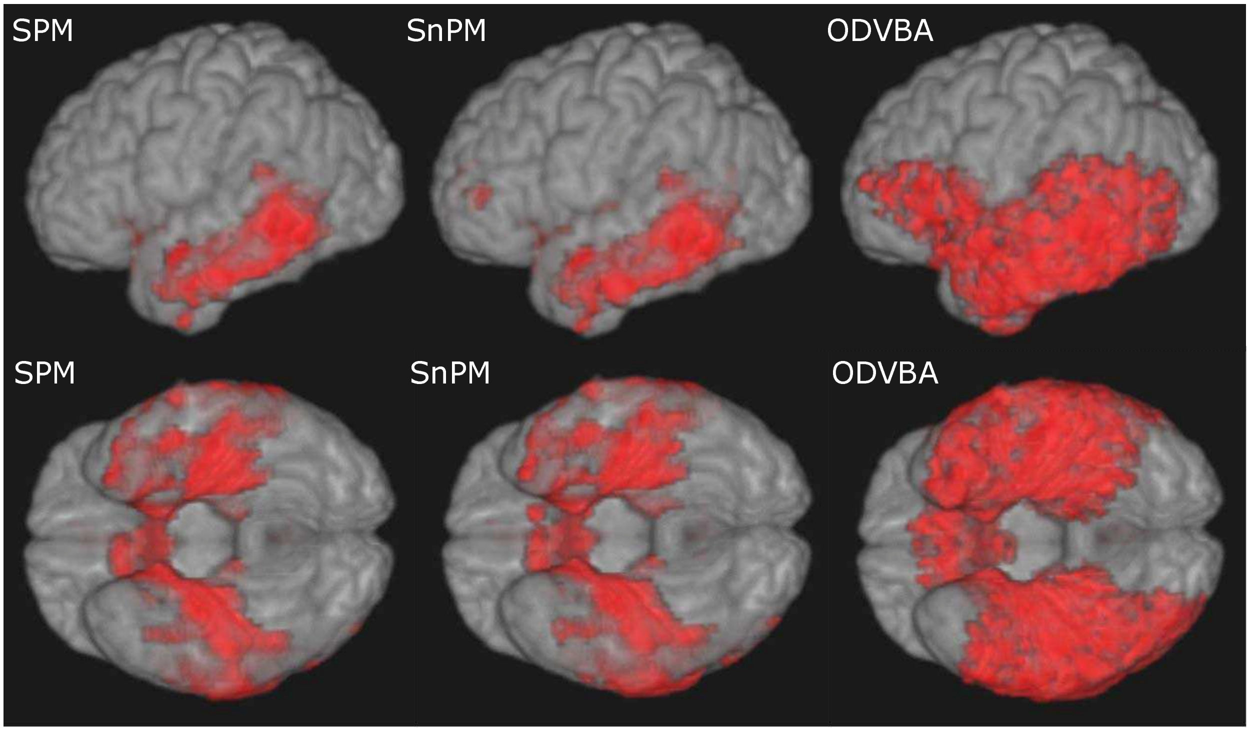

Fig. 13. Surface renderings of regions detected by the three methods.

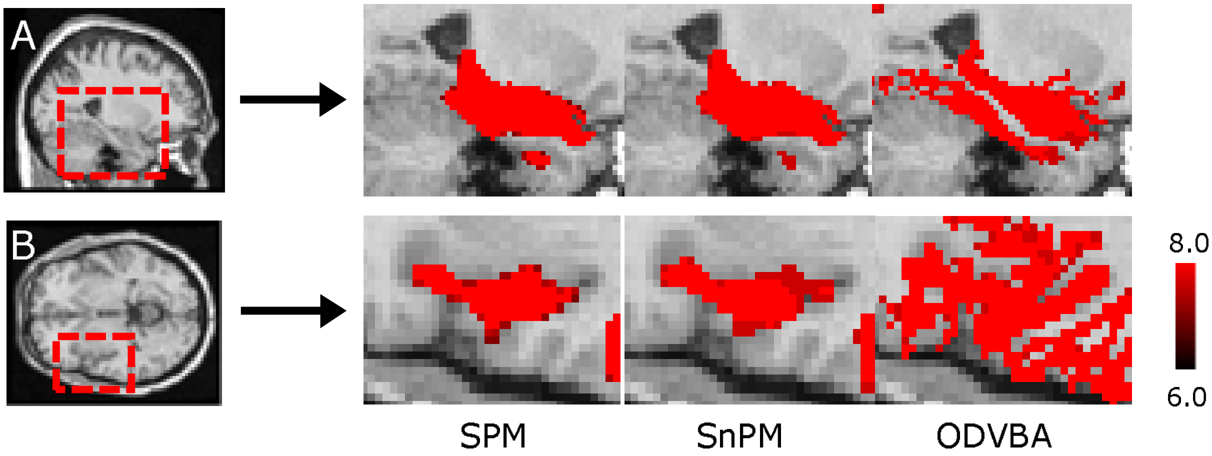

Fig. 14. Two representative magnified regions. The scale indicates the $-log(p)$ values.

but also identify many additional regions, e.g., Precuneus (Fig. 12A), Insula Cortex (Fig. 12B), Middle Temporal Gyrus (Fig. 12C), Lentiform (Fig. 12D), Occipital Lobe (Fig. 12E), Parahippocampal Gyrus (Fig. 12F), Superior Parietal Gyrus (Fig. 12G), Middle Frontal Gyrus (Fig. 12H), and Inferior Frontal Gyrus (Fig. 12I). As shown in the above mentioned sections in Fig. 12, the significant regions detected by ODVBA are either totally or partially missing from the results of SPM and SnPM. Fig. 13 shows the surface rendering of significant regions obtained from the three different methods. Moreover, these are all regions that are generally known from histopathology studies [4][5][15][21][22][34].

Fig. 14 shows two representative magnified regions that were discovered by SPM, SnPM and ODVBA. Among them, Fig. 14A shows the region near the Hippocampus and Fig. 14B shows the region around the Temporal Lobe. For the Fig. 14A, we can see that SPM and SnPM blurred the regions of the Hippocampus and Fusiform Gyrus. In contrast, a clear division between the two regions can be found in the results of ODVBA. For the Fig. 14B, SPM and SnPM blurred the different gyri and sulci in the region of the Temporal Lobe, while failing to detect other significant areas altogether; however, ODVBA delineates a more precise area of significant atrophy in that region.

We also employ the FDR procedure [25], a powerful approach commonly used in neuroimaging applications to

TABLE V
STATISTICS ON ANATOMICAL REGIONS.

Anatomical regions		p value<0.001						p value<0.05(corrected)					
		SPM		SnPM		ODVBA		SPM		SnPM		ODVBA	
		size	t	size	t	size	t	size	t	size	t	size	t
Middle Temporal Gyrus	L	1564	9.04	1686	8.74	2347	**9.25**	1309	9.00	1398	8.96	2244	**9.50**
	R	1779	8.86	2001	8.95	3058	**9.80**	1399	9.30	1663	9.07	2892	**9.95**
Superior Temporal Gyrus	L	1042	7.69	1131	7.74	1893	**10.14**	838	7.63	939	7.61	1785	**10.21**
	R	1046	7.68	1224	7.49	1899	**8.96**	841	7.77	1032	7.46	1830	**8.95**
Inferior Frontal Gyrus	L	194	6.07	271	7.08	882	**8.44**	171	6.26	179	6.14	763	**8.62**
	R	233	7.31	284	6.86	523	**7.86**	161	6.99	197	6.78	302	**7.79**
Superior Parietal Gyrus	L	298	6.70	335	6.60	571	**9.41**	328	6.57	300	6.70	462	**9.12**
	R	454	6.82	572	6.96	691	**8.64**	258	6.86	469	6.88	536	**8.58**
Precuneus	L	195	5.59	224	5.88	485	**8.35**	170	5.68	185	5.72	420	**8.09**
	R	282	6.02	304	5.97	415	**6.07**	220	6.06	191	**6.52**	362	5.91
Insula	L	595	6.61	624	6.62	838	**6.82**	470	7.41	491	7.12	748	**7.49**
	R	232	7.13	281	6.52	410	**7.24**	210	6.91	181	**7.44**	310	6.86
Middle Occipital Gyrus	L	6	2.37	42	4.76	183	**5.61**	0	\	0	\	128	**5.14**
	R	275	5.71	322	5.67	531	**7.43**	198	5.63	262	5.61	441	**7.62**
Hippocampus	L	360	6.59	373	6.53	330	**8.09**	345	6.81	365	6.52	319	**8.17**
	R	405	5.05	422	4.77	361	**7.03**	369	5.27	383	4.99	346	**7.29**

adjust for statistical error. FDR aims to control the portion of false positive error, instead of excluding this error [38]. In addition, FDR is adaptable since its value is computed directly on the observed p value distribution. We partitioned the resulting significant maps with p values of 0.001 and 0.05 (FDR corrected) respectively, according to predefined anatomical regions from the Jacob Atlas; calculated the means of the tissue density maps per region for all the samples; and finally, computed the t value based on these means according to the class information.

Table V lists the sizes and t statistics of major anatomical regions. We can see that not only the sizes detected by ODVBA, but also the corresponding t values are generally greater than those detected by SPM and SnPM. This means that the regions found by ODVBA display a higher degree of differentiation between the two groups and that SPM and SnPM might have missed some significant information.

2) Pattern Classification: In a separate experiment, we used the detected significant regions as the features input to a classifier of individual scans into NC or AD, and compared the three different methods in terms of the performance of classification with SVM. We randomly divided the original 50 (NC) +50 (AD) into 5 subsets (10+10 each), and then implemented 5-fold cross validation.

For each fold, 1) we used 40+40 for training and got the significant regions (with the p value $<$ 0.001 threshold) by the three different group analysis methods; 2) we calculated the means of the tissue density values of significant voxels in each predefined anatomical region (from the Jacob Atlas) for all the samples, and using the means from different regions as the input features, we got the SVM classifier [9]; 3) finally, we used the left testing set with size of 10+10 for validation.

The radial basis function (RBF) kernel is used in this study. Based on the 5-fold cross-validation, the classification rate of ODVBA (90%) is superior to that of SPM (86%) and that of SnPM (87%).

V. DISCUSSION AND CONCLUSION

In this paper, we have introduced a new framework, termed Optimally-Discriminative Voxel-Based Analysis (ODVBA), for group analysis of medical images. In the proposed framework, Nonnegative Discriminative Projection (NDP) is introduced to find the optimal discriminative direction in each learning set constructed by the neighborhood centered at the given voxel. Subsequently, each voxel's statistic is determined by a composition of all the smoothing directions which are associated with the given voxel. Finally permutation tests are used to obtain the statistical significance of the resulting ODVBA maps. In addition, to reduce the cost of computation, we developed a new method termed Graph-assisted Greedy Neighborhoods Cover to select a minimum subset of the neighborhoods which are used to learn the discriminative directions. We compared ODVBA to the traditional SPM and the nonparametric SPM (SnPM) with both simulated data and real AD data from the ADNI cohort. The experimental results have shown tested ODVBA against the conventional smoothing methods.

The main premise of our approach is that it effectively applies a form of matched filtering, to optimally detect a group difference. Since the shape of the target region of group difference is not known, regional discriminative analyses are used to identify voxels displaying the most significant differences. In addition to potentially improving sensitivity of detection of a structural or functional signal, this approach was shown in several experiments to better delineate the region of abnormality, in contrast with conventional smoothing approaches that blur through boundaries and dilute the signal from regions of interest with signal from regions that do not display a group difference.We use the coarse grid search over tuning parameters, e.g., ξ, ϕ, μ, and γ. Section IV-A.3) $Effect\ of\ kernel\ size$ and Section IV-A.4) $Effect\ of\ discrimination\ degree$ are examples of the parameter selection that we studied the performance of ODVBA on different parameters. Generally, our method is not so sensitive to the tuning parameters once its performance is stable. We can determine the optimal parameters in the simulated data in which we know the ground truth. If the simulated data and the real data come from the same data source, have similar sample size, and have similar size and degree of atrophy, we assume that the optimal parameters determined by the simulated data are also suitable for the real data. That is, the simulated data would be an appropriate reference for the real data to select the optimal parameters.

As described in the Introduction, our method are related with methods which use LDA and its variants over the entire image to determine "canonical image" that best discriminate between two groups or conditions. However our approach has significant differences from these approaches. In particular, such global LDA-based methods tend to be very limited by the small sample size and high-dimensionality problem, and therefore are typically able to only detect some of the regions that display group differences or activations, but not all. They are most suitable for classification purposes, rather than for voxel-based analysis. Most importantly, these methods produce both positive and negative loadings, which reflect cancelations in the data and therefore render it very difficult to interpret the data. In general, they are not designed to construct a precise spatial map of a group difference, but rather to find the best global discriminant. The non-negativity constraints in our ODVBA are essential. Our results can be interpreted in terms of brain activity or atrophy, for example, and the regions found match exactly the underlying regions of interest. The non-negativity constraints are by far not a trivial issue to implement, since they influence the entire optimization process.

APPENDIX A
PROOF OF THE POSITIVE DEFINITE MATRIX

If $\lambda_{min} \geq 0$, the smallest eigenvalue of $\gamma S_W - S_B + (|\lambda_{min}|+\tau^2)I$ is $2\lambda_{min}+\tau^2$ which is greater than 0. If $\lambda_{min} < 0$, the smallest eigenvalue of $\gamma S_W - S_B + (|\lambda_{min}| + \tau^2)I$ is just τ^2. In a word, all the eigenvalues of A is greater than 0. Since S_W, S_B, and I are all symmetric matrices, $A = \gamma S_W - S_B + (|\lambda_{min}| + \tau^2)I$ is a symmetric matrix. Thus, we complete the proof.

APPENDIX B
PROOF OF CONVERGENCE AND OPTIMALITY

Auxiliary function [32] is used to derive the rule of multiplicative updates in (10).

Definition 1. $G(\vec{v},\vec{w})$ is an auxiliary function for $J(\vec{w})$ if the conditions

$$\begin{cases} G(\vec{v},\vec{w}) \geq J(\vec{v}), \\ G(\vec{w},\vec{w}) = J(\vec{w}) \end{cases} \tag{22}$$

are satisfied.

Lemma 1. if $G(\vec{v},\vec{w})$ is an auxiliary function, then $J(\vec{w})$ is nonincreasing under the update:

$$\vec{w}' = \arg\min_{\vec{v}} G(\vec{v},\vec{w}). \tag{23}$$

One can refer to [32] for the proof of the above lemma. Following [42], we introduce such an auxiliary function for $J(\vec{w})$:

$$\begin{aligned} G(\vec{v},\vec{w}) =& \sum_i \frac{(A^+\vec{w})_i}{(\vec{w})_i} (\vec{v})_i^2 \\ & - \sum_{ij} A_{ij}^- (\vec{w})_i (\vec{w})_j \left(1 + log \frac{(\vec{v})_i (\vec{v})_j}{(\vec{w})_i (\vec{w})_j}\right) \\ & - \sum_i (\mu\vec{e})_i (\vec{v})_i. \end{aligned} \tag{24}$$

Note that $\sum_i \frac{(A^+\vec{w})_i}{(\vec{w})_i} (\vec{v})_i^2 \geq J_a(\vec{v})$ and $-\sum_{ij} A_{ij}^- (\vec{w})_i (\vec{w})_j \left(1 + log \frac{(\vec{v})_i (\vec{v})_j}{(\vec{w})_i (\vec{w})_j}\right) \geq -J_c(\vec{v})$ [42], so it is clear that the conditions $G(\vec{v},\vec{w}) \geq J(\vec{v})$ and $G(\vec{w},\vec{w}) = J(\vec{w})$ are met.

Rewrite the auxiliary function as the following form:

$$\begin{aligned} G(\vec{v},\vec{w}) =& \sum_i \frac{(A^+\vec{w})_i}{(\vec{w})_i} (\vec{v})_i^2 - 2\sum_i \left(A^-\vec{w}\right)_i (\vec{w})_i \, log \frac{(\vec{v})_i}{(\vec{w})_i} \\ & - \sum_i (\mu\vec{e})_i (\vec{v})_i - \sum_i \left(A^-\vec{w}\right)_i (\vec{w})_i \\ =& \sum_i G_i\left((\vec{v})_i\right) - \sum_i \left(A^-\vec{w}\right)_i (\vec{w})_i, \end{aligned} \tag{25}$$

where, $G_i\left((\vec{v})_i\right)$ means a function for each component in $\vec{v}$:

$$G_i\left((\vec{v})_i\right) = \frac{(A^+\vec{w})_i}{(\vec{w})_i} (\vec{v})_i^2 - 2\left(A^-\vec{w}\right)_i (\vec{w})_i \, log \frac{(\vec{v})_i}{(\vec{w})_i} - (\mu\vec{e})_i (\vec{v})_i . \tag{26}$$

Then, the minimization of $G(\vec{v},\vec{w})$ can be achieved by minimizing each $G_i\left((\vec{v})_i\right)$, and according to (23) we have:

$$(\vec{w}')_i = \arg\min_{(\vec{v})_i} G_i\left((\vec{v})_i\right). \tag{27}$$

Since the derivative of $G_i\left((\vec{v})_i\right)$ is:

$$G_i'\left((\vec{v})_i\right) = 2\frac{(A^+\vec{w})_i}{(\vec{w})_i} (\vec{v})_i - 2\left(A^-\vec{w}\right)_i \frac{(\vec{w})_i}{(\vec{v})\, i} - (\mu\vec{e})_i \,, \tag{28}$$

and considering $(\vec{w}')_i \geq 0$, we obtain the update rule as follows:

$$\begin{aligned} (\vec{w}')_i =& (\vec{v})_i|_{G_i'\left((\vec{v})_i\right)=0} \\ =& \left(\frac{(\mu\vec{e})_i + \sqrt{(\mu\vec{e})_i^2 + 16(A^+\vec{w})_i(A^-\vec{w})_i}}{4(A^+\vec{w})_i}\right) (\vec{w})_i . \end{aligned} \tag{29}$$

Acknowledgment

The authors would like to thank Alzheimer's Disease Neuroimaging Initiative (ADNI) for the data collection and sharing (supported by NIH grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the organizations as listed at www.adni-info.org. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation. They also thank Mr. Kayhan Batmanghelich for his contribution to the efficient software implementation of the algorithm.

References

[1] J. Ashburner and K. J. Friston, "Voxel-Based morphometry-the methods," *Neuroimage*, vol. 11, no. 6, pp. 805-821, 2000.

[2] J. Ashburner, J.G. Csernansky, C. Davatzikos, N.C. Fox, G.B. Frisoni, and P.M. Thomson, "Computer-assisted imaging to assess brain structure in healthy and diseased brains," *The Lancet Neurology*, vol. 2, no. 2, pp.79-88, 2003.

[3] J. C. Baron, G. Chetelat, B. Desgranges, G. Perchey, B. Landeau, S. V. de l and F. Eustache, "In vivo mapping of gray matter loss with voxel-based morphometry in mild Alzheimer's disease," *NeuroImage*, vol. 14, no.2, pp. 298-309, 2001.

[4] H. Braak, and E. Braak, "Neuropathological stageing of Alzheimer-related changes," *Acta Neuropathologica*, vol. 82, no. 4, 239-259, 1991.

[5] A. Brun and E. Englund, "Regional pattern of degeneration in Alzheimer's disease: neuronal loss and histopathological grading," *Histopathology*,vol. 5, pp. 548-564, 1981.

[6] T.A. Carlson, P. Schrater, and S. He, "Patterns of activity in the categorical representations of objects," *Journal of Cognitive Neuroscience*, vol. 15, no.5, pp. 704-717, 2003.

[7] R. Chen and E. Herskovits, "Graphical-model-based morphometric analysis," *IEEE Transactions on Medical Imaging*, vol. 24, no. 10, pp. 1237-1248, 2005.

[8] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2rd ed., Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

[9] R. Collobert and S. Bengio, "SVMTorch: Support Vector Machines for Large-Scale Regression Problems," *Journal of Machine Learning Research*, vol. 1, pp. 143-160, 2001.

[10] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein, *Introduction to Algorithms*, 2rd ed., MIT Press, 2001.

[11] A. M. Dale, A. K. Liu, B. R. Fischl, R. L. Buckner, J. W. Belliveau, J. D. Lewine, and E. Halgren, "Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity," *Neuron*, vol. 26, no. 1, pp. 55-67, 2000.

[12] C. Davatzikos, A. Genc, D. Xu, and S.M. Resnick, "Voxel-based morphometry using the RAVENS maps: Methods and validation using simulated longitudinal atrophy," *NeuroImage*, vol. 14, no. 6, pp. 1361-1369, 2001.

[13] C. Davatzikos, H.H. Li, E. Herskovits, and S.M. Resnick, "Accuracy and sensitivity of detection of activation foci in the brain via statistical parametric mapping: a study using a PET simulator," *NeuroImage*, vol. 13, no. 1, pp: 176-184, 2001.

[14] R. Duda, P. Hart and D. Stork, *Pattern Classification*, 2rd ed., Wiley, 2000.

[15] H. Engler, A. Forsberg, O. Almkvist, G. Blomquist, E. Larsson, I. Savitcheva, A. Wall, A. Ringheim, B. Langstrom, and A. Nordberg, "Two-year follow-up of amyloid deposition in patients with Alzheimer's disease," *Brain*, vol. 129, pp. 2856-2866, 2006.

[16] Y. Fan, D. Shen, R.C. Gur, R.E. Gur, and C. Davatzikos, "COMPARE: Classification Of Morphological Patterns using Adaptive Regional Elements," *IEEE Transaction on Medical Imaging*, vol. 26, no. 1, 93-105, 2007

[17] G.B. Frisoni, C. Testa, A. Zorzan, F. Sabattoli, A. Beltramello, H. Soininen, and M.P. Laakso, "Detection of grey matter loss in mild alzheimer's disease with voxel based morphometry," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 73, no.6, pp. 657-664, 2002.

[18] K.J. Friston, A.P. Holmes, K.J. Worsley, J.-B. Poline, C.D. Frith, and R.S.J. Frackowiak, "Statistical Parametric Maps in functional imaging: A general linear approach," *Human Brain Mapping*, vol. 2, pp. 189-210, 1995.

[19] A.F. Goldszal, C. Davatzikos, D. Pham, M. Yan, R.N. Bryan, and S.M. Resnick, "An image processing protocol for the analysis of MR images from an elderly population," *Journal of Computer Assisted Tomography*, vol. 22, no.5, pp. 827-837, 1998.

[20] J.V. Haxby, M.I. Gobbini, M.L. Furey, A. Ishai, J.L. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science*, vol. 293, pp. 2425-2430, 2001.

[21] J.A. Hardy and G.A. Higgins, "Alzheimer's disease: the amyloid cascade hypothesis," *Science*, vol. 256, pp. 184-185, 1992.

[22] K. Hensley, N.C. Hall, R. Subramaniam, P. Cole, M. Harris, M. Aksenov, M. Aksenova, S.P. Gabbita, J.F. Wu, J.M. Carney, M. Lovell, W.R. Markesbery, and D.A. Buttereld, "Brain regional correspondence between Alzheimer's disease histopathology and biomarkers of protein oxidation," *Journal of Neurochemistry*, vol. 65, 2146-2156, 1995.

[23] E. Herskovits, H. Peng, C. Davatzikos, "A Bayesian morphometry algorithm," *IEEE Transactions on Medical Imaging*, vol. 23, no. 6, pp. 723-737, 2004

[24] A.P. Holmes, R.C. Blair, J.D. Watson, and I. Ford, "Nonparametric analysis of statistic images from functional mapping experiments," *Journal of Cerebral Blood Flow and Metabolism*, vol. 16, no. 1, pp. 7-22, 1996.

[25] C.R. Genovese, N.A. Lazar, T. Nichols, "Thresholding of statistical maps in functional neuroimaging using the false discovery rate," *NeuroImage*, no. 15, pp. 870-878, 2002.

[26] N. Kabani, D. MacDonald, C.J. Holmes, A. Evans, "A 3D atlas of the human brain," *NeuroImage*, vol. 7, no.4, pp. S717, 1998.

[27] G.B. Karas, E.J. Burton and S.A. Rombouts, R.A. van Schijndel, J.T. O'Brien, P.h. Scheltens, I.G. McKeith, D. Williams, C. Ballard and F. Barkhof, "A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry," *Neuroimage*, vol. 18, no. 4, pp. 895-907, 2003.

[28] S. Kloppel, C.M. Stonnington, C. Chu, B. Draganski, R.I. Scahill, J.D. Rohrer, N.C. Fox, C.R. Jack Jr, J. Ashburner, and R.S.J. Frackowiak "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no.3, pp. 681-689, 2008.

[29] N. Kriegeskorte, R. Goebel, and P. Bandettini, "Information-based functional brain mapping," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 3863-3868, 2006.

[30] N. Kriegeskorte, P. Bandettini, "Analyzing for information, not activation, to exploit high-resolution fMRI," *NeuroImage*, vol. 38, no. 4649-662, 2007.

[31] R. Kustra, S.C. Strother, "Penalized discriminant analysis of $\left[{}^{15}O\right]$ water PET brain images with prediction error selection of smoothing and regularization hyperparameters," *IEEE Transaction on Medical Imaging*, vol. 20, pp. 376-387, 2001.

[32] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," In *Advances in Neural Information Processing*, vol. 13, 2001.

[33] T. Matthew, W.M. Wells III, C.D. Louis, and T. Arbel, "Feature-Based Morphometry: Discovering Group-related Anatomical Patterns," *NeuroImage*, vol. 49, no. 3, pp. 2318-2327, 2010.

[34] A.C. McKee, K.S. Kosik, and N.W. Kowall, "Neuritic pathology and dementia in Alzheimer's disease," *Annals of Neurology*, vol. 30, pp. 156-165, 1991.

[35] C.E. Metz, "ROC methodology in radiologic imaging," *Investigative Radiology*, vol. 21, pp. 720-733, 1986.

[36] C. Misra, Y. Fan, and C. Davatzikos, "Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI," *Neuroimage*, vol. 44, no. 4, pp. 1415-1422, 2009.

[37] J.H. Morra, Z. Tu, L.G. Apostolova, A.E. Green, A.W. Toga, P.M. Thompson, "Comparison of adaBoost and support vector machines for detecting alzheimer's disease through automated hippocampal segmentation," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 30-43, 2010.

[38] T. Nichols, S. Hayasaka, "Controlling the familywise error rate in functional neuroimaging: a comparative review," *Statistical methods in medical research*, vol. 12, no. 5, pp. 419-446, 2003.

[39] T.E. Nichols and A.P. Holmes, "Nonparametric permutation tests for functional neuroimaging: a primer with examples," *Human Brain Mapping*, vol. 15, no. 1, pp: 1-25, 2002.

[40] K.A. Norman, S.M. Polyn, G.J. Detre, and J.V. Haxby, "Beyond mind-reading: Multi-voxel pattern analysis of fMRI data," *Trends in Cognitive Sciences*, vol. 10, no. 9, pp. 424-430, 2006.

[41] D. L. Pham and J. L. Prince, "Adaptive fuzzy segmentation of magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 9, pp. 737-752. 1999.

[42] F. Sha, Y. Lin, L.K. Saul, and D.D. Lee, "Multiplicative updates for nonnegative quadratic programming," *Neural Computation*, vol. 19, no. 8, pp. 2004-2031, 2007.

[43] D. Shen and C. Davatzikos, "HAMMER: Hierarchical attribute matching mechanism for elastic registration," *IEEE Transactions on Medical Imaging*, vol. 21, no. 11, pp. 1421-1439, 2002.

[44] SPM, Available: www.fil.ion.ucl.ac.uk/spm/

[45] C. E. Thomaz, J. P. Boardman, S. Counsell, D. L. G. Hill, J. V. Hajnal, A. D. Edwards, M. A. Rutherford, D. F. Gillies and D. Rueckert, "A multivariate statistical analysis of the developing human brain in preterm infants," *Image and Vision Computing*, vol. 25, no.6, pp: 981-994, 2007.

[46] C. E. Thomaz, F. Duran, G. F. Busatto, D. F. Gillies and D. Rueckert, "Multivariate statistical differences of MRI samples of the human brain," *Journal of Mathematical Imaging and Vision*, vol. 29, no.2-3, pp:95-106, 2007.

[47] P.M. Thompson, K.M. Hayashi, G. de Zubicaray, A.L. Janke, S.E. Rose, J. Semple, D. Herman, M.S. Hong, S.S. Dittmer, D.M. Doddrell and A.W. Toga, "Dynamics of gray matter loss in Alzheimer's disease," *Journal of Neuroscience*, vol. 23, no. 3, pp. 994-1005, 2003.

[48] D. Van De Ville, M.L. Seghier, F. Lazeyras, T. Blu, M. Unser, "WSPM: Wavelet-Based Statistical Parametric Mapping," *Neuroimage*, vol. 37, no. 4, pp. 1205-1217, 2007.

[49] A.M. Wink and J.B.T.M. Roerdink, "Denoising functional MR Images: a comparison of wavelet denoising and Gaussian smoothing," *IEEE Transactions on Medical Imaging*, vol. 23 no. 3, pp. 374-387, 2004.

[50] L. Xu, G. Pearlson, and V. Calhoun, "Source Based Morphometry: The Use of Independent Component Analysis to Identify Gray Matter Differences with Application to Schizophrenia," *Human Brain Mapping*, vol. 30, pp. 711-724, 2009.

[51] L. Yang, "Alignment of Overlapping Locally Scaled Patches for Multi-dimensional Scaling and Dimensionality Reduction," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp: 438-450, 2008.