Optimally-Discriminative Voxel-Based Analysis

Tianhao Zhang and Christos Davatzikos

Section of Biomedical Image Analysis, Department of Radiology,
University of Pennsylvania, Philadelphia, PA 19104, USA
{Tianhao.Zhang,Christos.Davatzikos}@uphs.upenn.edu

Abstract. Gaussian smoothing of images is an important step in Voxel-based Analysis and Statistical Parametric Mapping (VBA-SPM); it accounts for registration errors and integrates imaging signals from a region around each voxel being analyzed. However, it has also become a limitation of VBA-SPM based methods, since it is often chosen empirically, non-optimally, and lacks spatial adaptivity to the shape and spatial extent of the region of interest. 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 locally 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. Permutation tests are finally used to obtain the statistical significance. The experiments on Mild Cognitive Impairment (MCI) study have shown the effectiveness of the framework.

1 Introduction

Voxel-based Analysis and Statistical Parametric Mapping (VBA-SPM) [2][7] 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 interests (ROIs). A fundamentally important aspect of VBA-SPM has been the spatial smoothing of images prior to analysis. Typically, Gaussian blurs of full-width-half-max (FWHM) in the range of 8-15mm are used to account for registration errors, to Gaussianize 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 confound the analysis; 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

T. Jiang et al. (Eds.): MICCAI 2010, Part II, LNCS 6362, pp. 257–265, 2010.

functional activation or of structural abnormalities. 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.

However, the most profound limitation of Gaussian smoothing of images is its lack of spatial adaptivity to the shape and spatial extent of the 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, and the white matter. Some earlier work in the literature [5] had shown that spatially adaptive filtering of image data can improve statistical power to detect group differences, however it didn't offer a way to determine optimal data filtering. In general, 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 medical 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. In order to determine the optimal smoothing kernel, a local discriminative analysis, restricted by appropriate nonnegativity constraints, is 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.

2 The Proposed Framework

The proposed framework contains three stages: 1) Local Nonnegative Discriminative Projection, 2) Determining each voxel's statistic, and 3) Permutation tests.

2.1 Local Nonnegative Discriminative Projection

Learning Set Construction. For a given voxel x in volume X, we construct its neighborhood $\mathbb{N}$: $\|x - x_i\| < \xi$. To render subsquent processing tractable, 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: $\boldsymbol{\theta} = [x, x_1, \cdots, x_{k-1}]^T$. Provided that there are N subjects, we can obtain N subvolume vectors which form a data set: $\Theta = [\theta_1, \theta_2, \cdots, \theta_N]$ for learning. The procedure is illustrated in Fig. 1.

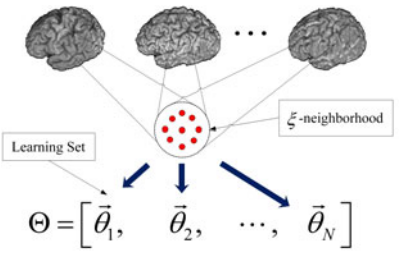

Fig. 1. Learning set construction

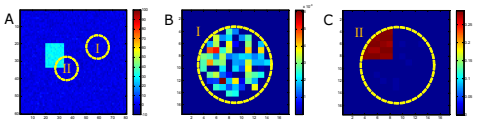

Fig. 2. Illustration of the basic idea of NDP using a toy dataset

The Basic Idea of NDP. The Nonnegative Discriminative Projection (NDP) algorithm is used to find the optimal discriminative directions which project the high-dimensional subvolume samples onto a 1-dimensional space maximizing the classification accuracy. The resultant optimally filter $\boldsymbol{w}$ is nonnegative, because of the nonnegativity constraint incorporated into the objective function. This constraint is used to help us interpret the group differences. Specifically, our goal is not simply to find an image contrast, prescribed by $\boldsymbol{w}$, 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, and therefore can be easily interpreted.

To illustrate the idea of NDP, we show its results on a toy dataset before describing the formulation. We generated two groups of images containing a square with intensity varying from one image to another: 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 $\boldsymbol{w}$ obtained from the learning set constructed according to the neighborhood I; it is basically noise with very small values of $(\boldsymbol{w})_j$, indicating that no local filter can be found that distinguishes the two groups at that neighborhood. Fig. 2C shows the $\boldsymbol{w}$ obtained from the learning set corresponding to neighborhood II; the estimated $\boldsymbol{w}$ is well aligned with the underlying group difference, within which it has high values. The bottom-line here is that a properly estimated $\boldsymbol{w}$ can highlight the underlying difference.

The Formulation of NDP. Using one given learning set, we probe into neighborhood elements' contributions for discrimination of the two groups. We target

to find such a nonnegative vector $\boldsymbol{w}$: the larger the value of $(\boldsymbol{w})_j$ is, the more the corresponding element $(\boldsymbol{\theta})_j$ contributes to the discrimination. Equivalently, $(\boldsymbol{w})_j$ is the jth coefficient of the regional filter denoted by $\boldsymbol{w}$. Via $\boldsymbol{w}$, the learning set can be projected from the k-dimensional space onto the 1-dimensional space to be optimally classified, such as $\Psi = \boldsymbol{w}^T\boldsymbol{\theta}$. We expect that, in the projected space, the two classes will be separated as much as possible along $\boldsymbol{w}$, and at the same time the samples from the same class get more compact. A measure of the separation between the two classes is $\boldsymbol{w}^T S_B \boldsymbol{w}$, where $S_B = (\boldsymbol{m_1} - \boldsymbol{m_2})(\boldsymbol{m_1} - \boldsymbol{m_2})^T$; $\boldsymbol{m_i} = \frac{1}{N_i}\sum_{\boldsymbol{\theta}\in C_i}\boldsymbol{\theta}$; C_i means the ith class; N_i denotes the number of samples in C_i. And, the intraclass compactness can be described by $\boldsymbol{w}^T S_W \boldsymbol{w}$, where $S_W = \sum_{i=1}^{2}\sum_{\theta\in C_i}(\boldsymbol{\theta} - \boldsymbol{m_i})(\boldsymbol{\theta} - \boldsymbol{m_i})^T$. S_B and S_W are called the between-class scatter matrix and the within-class scatter matrix respectively, according to the classic Fisher LDA [6] in which the criterion function is based on the generalized Rayleigh quotient. Herein, S_B and S_W are considered under the formulation of quadratic programming which is amenable to the nonnegative constraint as follows:

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

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

Theorem 1. *A is a positive definite matrix.*

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

Since A is positive definite, $J(\boldsymbol{w})$ is a convex function and has the unique global minimum. We solve the above optimization problem using the Nonnegative Quadratic Programming (NQP) [11]. According to [11], define the nonnegative matrices A^+ and A^- as follows: $A^+_{ij} = A_{ij}$, if $A_{ij} > 0$; otherwise, it is 0. $A^-_{ij} = |A_{ij}|$, if $A_{ij} < 0$; otherwise it is 0. So it is clear that $A = A^+ - A^-$.

Multiplicative updates rule which does not involve the learning rates, is introduced to minimize the objective function iteratively:

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

where $i = 1, \cdots, k$. Eq.2 means that all the elements in $\boldsymbol{w}$ are updated in parallel. Since $(A^+\boldsymbol{w})_i \geq 0$ and $(A^-\boldsymbol{w})_i \geq 0$, the updated $\boldsymbol{w}$ in Eq.2 is always nonnegative.

Theorem 2. *The function of $J(\boldsymbol{w})$ in Eq.1 decreases monotonically to the value of its global minimum under the multiplicative updates in Eq.2.*

proof. An auxiliary function ([9], pp. 659, Definition 1)([11], pp. 2013, Theorem 1) as follows is used to derive the multiple updates:

$$G(\boldsymbol{v},\boldsymbol{w}) = \sum_i \frac{(A^+\boldsymbol{w})_i}{(\boldsymbol{w})_i}(\boldsymbol{v})_i^2 - \sum_{ij} A_{ij}^-(\boldsymbol{w})_i(\boldsymbol{w})_j\left(1 + log\frac{(\boldsymbol{v})_i(\boldsymbol{v})_j}{(\boldsymbol{w})_i(\boldsymbol{w})_j}\right) - \sum_i (\mu\boldsymbol{e}^T)_i(\boldsymbol{v})_i. \tag{3}$$

According to ([9], pp. 659, Lemma 1), $J(\boldsymbol{w})$ is nonincreasing under the updates: $\boldsymbol{w} = \arg\min_{\boldsymbol{v}} G(\boldsymbol{v},\boldsymbol{w})$ and for each component in $\boldsymbol{w}$, $(\boldsymbol{w})_i = (\boldsymbol{v})_i|_{G_i'=0}$, where $G_i' = 2(A^+\boldsymbol{w})_i(\boldsymbol{v})_i/(\boldsymbol{w})_i - 2(A^-\boldsymbol{w})_i(\boldsymbol{w})_i/(\boldsymbol{v})_i - (\mu\boldsymbol{e}^T)_i$. So, we can obtain the updates described as Eq. 2. □

2.2 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 2.1. For a given voxel x, we obtain a list of $(\boldsymbol{w})_j$ values since x may belong to a number of neighborhoods. To quantify the group difference measured at voxel x , we use the *discrimination degree*, which relates to the effect size [3]:

$$\delta = \left(\frac{|\tilde{m}_1 - \tilde{m}_2|}{\sqrt{\sum_{i=1}^{2}\sum_{\Psi\in C_i}(\Psi - \tilde{m}_i)^2}}\sqrt{N_1+N_2-2}\right)^{\phi}, \tag{4}$$

where, $\tilde{m}_i = \frac{1}{N_i}\sum_{\Psi\in C_i}\Psi$, ϕ is the tuning parameter for reducing potential outliers in the dataset. Let $\Delta = \{\mathbb{N}|x \in \mathbb{N}\}$ denote the set of neighborhoods that a voxel x belongs to, then we define the group difference on x by summing up contributions from all neighborhoods to which it participates:

$$S_x = \sum_{\mathbb{N}\in\Delta} \delta_{\mathbb{N}}\left|(\boldsymbol{w}_{\mathbb{N}})_j\right|,\ j \in \{1,\cdots,k\}, \tag{5}$$

where, $\boldsymbol{w}_{\mathbb{N}}$ denotes the coefficients corresponding to voxels in $\mathbb{N}$, $(\boldsymbol{w}_{\mathbb{N}})_j$ denotes that x is the jth element in $\mathbb{N}$, and $\delta_{\mathbb{N}}$ which acts as the weight for $\boldsymbol{w}_{\mathbb{N}}$ denotes the *discrimination degree* achieved in neighborhood $\mathbb{N}$ and is defined in Eq. 4. S_x will serve as the statistic reflecting group differences on the voxel x, and will be used next to determine statistical significance. Higher values of S_x reflect stronger group differences.

2.3 Permutation Tests

Assume the null hypothesis that 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 [10]. In particular, we randomly assign the subjects into two groups, and then implement Section 2.1- Section 2.2 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 $S_i, i = 1, \cdots, N_P$ denotes the ones obtained by relabeling. The P value for the given voxel is calculated according to:

$$P = \sum_{i=1}^{N_P} \left[u(S_i - S_0)\right] / N_P \tag{6}$$

where, $u(t) = 1$, if $t \geq 0$; otherwise it is 0.

3 Results

In this section, we carry out the study of determining the extent of atrophy in Mild Cognitive Impairment (MCI) subjects to evaluate ODVBA compared with the original SPM [2] and the nonparametric permutation based SPM (SnPM) [10]. The data was obtained from ADNI [1], which has recruited approximately 800 adults including 200 normal controls, 400 individuals with MCI and 200 Alzheimer's disease (AD) patients. The images were acquired and processed according to a number of steps detailed under the ADNI website [1]. We randomly selected 100 subjects with MCI from the ADNI cohort. 50 of these subjects that had undergone conversion to AD were referred to MCI-C. The remaining 50 non-converters were referred to MCI-NC. Images were preprocessed according to the following steps. 1) Alignment to the ACPC plane; 2) Removal of extracranial material; 3) Tissue segmentation into grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF); 4) High-dimensional image warping to a standardized coordinate system; 5) Formation of tissue density maps typically used in the modulated SPM analysis [4]. We used GM for evaluation purposes.

Both SnPM and ODVBA are implemented with 2000 permutations. Fig. 3 shows some selected sections from the results (with P value < 0.001 thresthold) of

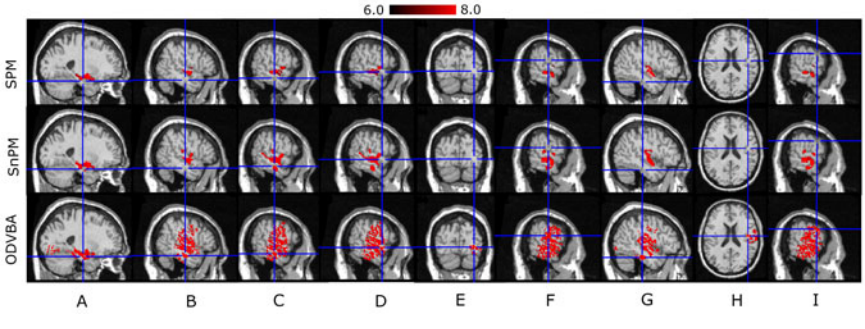

Fig. 3. Representative sections with significant regions. The scale indicates the $-log(P)$ values.

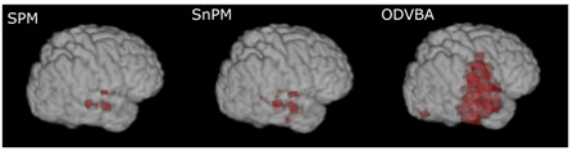

Fig. 4. Surface renderings of regions

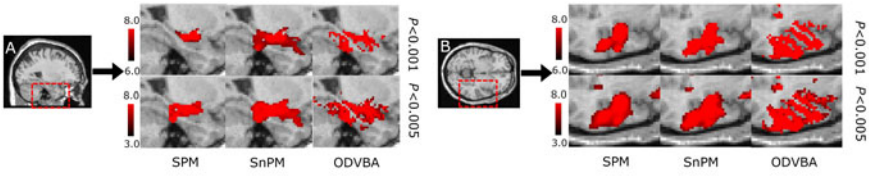

Fig. 5. Representative magnified regions with two levels of P value

SPM, SnPM, and ODVBA, respectively. We can see that the results of ODVBA reflect significant GM loss in MCI-C compared with MCI-NC mainly in Hippocampus (Fig. 3A), Inferior Temporal Gyrus (Fig. 3B), Middle Temporal Gyrus (Fig. 3C), Superior Temporal Gyrus (Fig. 3D), Occipital Lobe (Fig. 3E), Insular Cortex (Fig. 3F), Fusiform Gyrus (Fig. 3G), Parietal Lobe (Fig. 3H), and Inferior Frontal Gyrus (Fig. 3I). In addition, some significant regions detected by ODVBA are either totally or partially obscured in the results of SPM and SnPM. Surface renderings in Fig. 4 also demonstrate the results of the three methods. As we can see, ODVBA reveals a substantially larger and more detailed area of atrophy in Temporal Lobe, Parietal Lobe, and Frontal Lobe than SPM and SnPM, which barely detected the damage. Moreover, these are all regions that are generally known from histopathology studies to be affected in AD.

Fig. 5 shows two representative magnified regions that were found by the three methods. Among them, Fig. 5A shows the region near the Hippocampus and Fig. 5B shows the region around the Temporal Lobe. The results are with P value < 0.005 threshold and P value < 0.001 threshold respectively. For the Fig. 5A, we can see that SPM and SnPM blurred the regions of the Hippocampus

Table 1. Statistics of Clusters

Cluster	P-value<0.005						P-value<0.001					
	SPM		SnPM		ODVBA		SPM		SnPM		ODVBA	
	size	t	size	t	size	t	size	t	size	t	size	t
#1	10657	7.86	13463	7.65	13816	**8.47**	2430	8.12	2430	7.35	9568	**8.70**
#2	398	4.96	737	4.79	2341	**8.18**	1424	6.81	1424	6.34	855	**7.09**
#3	325	5.33	359	5.43	1657	**6.93**	1365	5.27	1365	6.99	575	**7.28**

and the Fusiform Gyrus. In contrast, a clear separation of the two regions can be found in the results of ODVBA. Moreover, the results of SPM with P value < 0.001 detected no significant atrophy located in the Fusiform Gyrus in that section. For the Fig. 5B, SPM and SnPM blurred the different gyri and sulci in the region of the Temporal Lobe; however, ODVBA delineates a more precise area of significant atrophy in that region.

In Table 1, we list the t value on the three biggest clusters with P values < 0.005 and < 0.001, respectively. The t value is calculated based on the cluster means on the tissue density maps of the two groups' samples. We can see that the t values with clusters of ODVBA are higher than those of SPM and SnPM; that is, the regions found by ODVBA display a greater difference between the two groups than SPM and SnPM.

4 Summary

We have introduced a new framework of voxel-based analysis, aiming to detect differences associated with brain abnormalities existing in two groups. The main premise of this approach is that the optimal shape and size of the spatial filter to be applied to the data prior to statistical analysis is not known in advance, but must be estimated from the data. Moreover, this spatial filtering is not fixed throughout the image, as customary in the literature, but is spatially adaptive, depending on the local anatomy and abnormality (which is unknown in advance, as well). We presented a nonnegative discriminative direction method, which determines the filter that best distinguishes the two groups being compared. This approach was evaluated in the study of contrasting MCI-C versus MCI-NC, and revealed substantially more extensive, and more significant GM atrophy in regions known to be affected by AD, whereas SPM and SnPM produced inferior results.

References

1. Alzheimer's Disease Neuroimaging Initiative, `http://www.loni.ucla.edu/ADNI`
2. Ashburner, J., Friston, K.J.: Voxel-based morphometry-the methods. Neuroimage 11(6), 805–821 (2000)
3. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences, 2nd edn. Lawrence Erlbaum Associates, Mahwah (1988)
4. 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)
5. Davatzikos, C., Li, H.H., Herskovits, E., Resnick, S.M.: Accuracy and sensitivity of detection of activation foci in the brain via statistical parametric mapping: a study using a PET simulator. NeuroImage 13(1), 176–184 (2001)
6. Duda, R., Hart, P., Stork, D.: Pattern Classification, 2nd edn. Wiley, Chichester (2000)

7. Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.B., Frith, C.D., Frackowiak, R.S.J.: Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Map. 2, 189–210 (1995)
8. Genovese, C.R., Lazar, N.A., Nichols, T.: Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15, 870–878 (2002)
9. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: NIPS, vol. 13, pp. 556–562 (2001)
10. Nichols, T.E., Holmes, A.P.: Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Map. 15(1), 1–25 (2002)
11. Sha, F., Lin, Y., Saul, L.K., Lee, D.D.: Multiplicative updates for nonnegative quadratic programming. Neural. Comp. 19(8), 2004–2031 (2007)