Sparse classification for significant anatomy detection in a group study

Dana Cobzas
University of Alberta

with
Linglong Kong Math, U of Alberta
Mark Schmidt CS, U of British Columbia
Alan Wilman BME, U of Alberta
What are the areas of the brain significantly different between healthy subjects and patients?
Voxel based analysis – data alignment

Scalar images

Subject 1  Subject 2  [ ... ]  Subject n

Linear and nonlinear registration in atlas space
Voxel based analysis – data alignment

Scalar images

Subject 1  Subject 2  Subject n

[...]  Linear and nonlinear registration in atlas space

Flattened data: Only relevant parts

Data
\[ p \text{ voxels} \]

Labels
Control/Patient

\[ A_{n\times p} \]
Voxel based analysis:
• Voxel-based univariate approaches, permutations tests, FDR
• Advanced statistics to compensate for data correlation
VBA – traditional approaches

Voxel based analysis:
• Voxel-based univariate approaches, permutations tests, FDR
• Advanced statistics to compensate for data correlation

Reformulate the problem as a supervised dimensionality data reduction method such that
• The detected anatomy is discriminative
• Sparse
• Compact and interpretable from an anatomical viewpoint
• There is a principled way of finding an optimal model and testing its accuracy
Dimensionality reduction

Unsupervised:
• PCA, ICA: eigenvectors have global support and do not provide anatomical specificity
Dimensionality reduction

Unsupervised:

- PCA, ICA: eigenvectors have global support and do not provide anatomical specificity \(\Rightarrow\) Imposed sparseness of solution

- Sparse generative models “parts-based representations”;
  But they do not explicitly optimize discrimination

[Kandel, Avants et al. 2015][Lee, Seung 1999][Witten, Hastie 2009]
Dimensionality reduction

**Unsupervised:**
- PCA, ICA: eigenvectors have global support and do not provide anatomical specificity → Imposed **sparseness** of solution
- Sparse generative models “parts-based representations”;
  But they do not explicitly optimize discrimination
  [Kandel, Avants et al. 2015][Lee, Seung 1999][Witten, Hastie 2009]

**Supervised:**
- Pattern classification methods – feature extraction and selection to achieve high classification accuracy; sparsity on feature selection;
- Goal: high classification accuracy with no focus on meaning and interpretability of selected regions; often data reduction decoupled from feature selection
  [Batmanghelich et al. 2011][Sabuncu, Van Leemput 2012]
  [Krishnapuram et al. 2005][Ryali et al. 2010]
Sparse classification

Discriminative method where relevant image regions are selected using an image-regularized sparse classification.

- The detected anatomy is discriminative
- Sparse
- Compact and interpretable from an anatomical viewpoint
- There is a principled way of finding an optimal model and testing its accuracy

- Logistic regression
- Sparseness constraint
- Image-based regularization
- Cross-validation

Similar to [Kandel et al. 2013] work on sparse regression
Formulation: Logistic regression

Data
p voxels

Labels -1/1

\( y_i \) follows a logistic regression distribution with location \( a_i x + b \)

Image coefficients \( x \)
Scalar bias \( b \)

Optimal params \( x, b \)
Formulation: sparseness

Looking for a \textbf{sparse} solution $x$

$$\min_{x,b} \sum_{i=1}^{n} \log(1 + \exp(-y_i(a_i x + b)))$$

subject to $\|x\|_0 \leq s$
Formulation: sparseness

Looking for a **sparse** solution $x$

$$\min_{x,b} \sum_{i=1}^{n} \log (1 + \exp(-y_i(a_i x + b)))$$

subject to $\|x\|_0 \leq s$

NP hard
Replace $l_0$ with $l_1$
The two penalties give identical solution for many problems

$$\min_{x,b} \sum_{i=1}^{n} \log (1 + \exp(-y_i(a_i x + b))) + \lambda_1 \|x\|_1$$

Non smooth at zero catches many solutions
Formulation: sparseness

Looking for a **sparse** solution $x$

$$\min_{x,b} \sum_{i=1}^{n} \log (1 + \exp(-y_i(a_i x + b)))$$

subject to $\|x\|_0 \leq s$

$$\min_{x,b} \sum_{i=1}^{n} \log (1 + \exp(-y_i(a_i x + b))) + \lambda_1 \|x\|_1$$

NP hard
Replace $l_0$ with $l_1$
The two penalties give identical solution for many problems

Non smooth at zero catches many solutions

**L2-regularization**

**L1-regularization**
Formulation: compactness

Problem:
- We like the results to have an anatomical interpretation
- Add image-based regularization terms
Formulation: compactness

Problem:
- We like the results to have an anatomical interpretation
- Add image-based regularization terms

Diffusion based regularization

- Uniform diffusion

- Total variation TV-$l_1$
  Nonlinear isotropic diffusion [Weickert]

\[
\begin{align*}
\text{Diffusion} & \quad \partial_t x = \Delta x \\
\text{Regularization} & \quad ||\nabla x||_2^2 \\
& \quad ||\nabla x||_{2,1} = \sqrt{\partial_x x^2 + \partial_y x^2 + \partial_z x^2} \\
\end{align*}
\]

\[
\begin{align*}
\partial_t x & = \Psi'(||\nabla x||^2) \\
\Psi(||\nabla x||^2) & = \sqrt{\beta^2 + s^2} \\
\Psi'(s) & = \frac{1}{\sqrt{\beta^2 + s^2}}
\end{align*}
\]
Formulation: compactness

Problem:
- We like the results to have an anatomical interpretation
- Add image-based regularization terms

Diffusion based regularization
- Uniform diffusion
- Total variation TV-$l_1$
  Nonlinear isotropic diffusion [Weickert]

\[ \partial_t x = \Delta x \]
\[ \partial_t x = \Psi'(||\nabla x||^2) \]
\[ \Psi(s) = \sqrt{s} \]
\[ \Psi'(s) = \frac{1}{2\sqrt{s}} \]
Optimization

\[
\min_{x,b} \sum_{i=1}^{n} \log (1 + \exp(-y_i(a_i x + b))) + \lambda_1 \|x\|_1 + \lambda_2 \|\nabla x\|_2^2
\]

Solving this optimization is complicated due to non-differentiability of \(l_1\) terms. Use projected gradient methods

- choose an approximate steepest descent direction (pseudo-gradient) of the objective function
- take an approximate Newton step
- project solution (make \(x_k=0\) if sign changed)

\[x^+ = \text{project}_C [x - \alpha f'(x)],\]
**Experiments**

Compare 3 methods:

- **Eigenanatomy** [Avants et al.]
  - unsupervised detection of sparse regions using sparse PCA
  - use projection of data onto eigenvectors in a logistic regression classification method

- **SurfStat** [Worsley et al.]
  - compute significant regions from using RFT implemented in SurfStat
  - use detected significant voxels in a logistic regression classification

- **SparseClassification**
  - Each method returns – classification labels \( \tilde{y} \) and sparse regions \( \tilde{x} \) that are compared with ground truth \( y \) and \( x \).

For each data \( A \) we computed classification results using **3 folds cross-validation** >
data divided in 3 groups, 2 used for training and one for testing – all 3 combinations
Optimal params for each method are determined using cross-validation for each dataset \( A \). Only results with optimal params are considered.

**Reported measures**

- Sparseness of detected regions \( sp \)
- Classification results (\( y \) vs \( \tilde{y} \)) : accuracy, sensitivity, specificity, AUC
- Accuracy and stability of regions (\( x \) vs \( \tilde{x} \)):
  - dice score ; dice overlap for the 3 folds \( \tilde{x}_1, \tilde{x}_2, \tilde{x}_3 \)
- Significance of regions in a t-test : use mean data in sparse regions \( x \) in a t-test \( p\text{-val} \)
Compare **3 methods**:  

- **Eigenanatomy** [Avants et al.]  
  unsupervised detection of sparse regions using sparse PCA  
  use projection of data onto eigenvectors in a logistic regression classification method  

- **SurfStat** [Worsley et al.]  
  compute significant regions from using RFT implemented in SurfStat  
  use detected significant voxels in a logistic regression classification  

- **SparseClassification**  
  Each method returns – classification labels \( \tilde{y} \) and sparse regions \( \tilde{x} \) that are compared with ground truth \( y \) and \( x \).

For each data \( A \) we computed classification results using **3 folds cross-validation** > data divided in 3 groups, 2 used for training and one for testing – all 3 combinations  
Optimal params for each method are determined using cross-validation for each dataset \( A \). Only results with optimal params are considered.

**Reported measures**  
- Sparseness of detected regions \( sp \)  
- Classification results (\( y \) vs \( \tilde{y} \)) : accuracy, sensitivity, specificity, AUC  
- Accuracy and stability of regions (\( x \) vs \( \tilde{x} \)) : dice score, dice overlap for the 3 folds \( \tilde{x}_1, \tilde{x}_2, \tilde{x}_3 \)  
- Significance of regions in a t-test : use mean data in sparse regions \( \tilde{x} \) in a t-test \( p\text{-val} \)  

**Dice Score**:

\[
\frac{2(A \cap B)}{|A| + |B|}
\]
Synthetic data

\[ \mathbf{a}_i: \]
32 x 32 x 8

\[ > \mathbf{a}_i \sim \mathcal{N}(0,1) \]

> 4 blocks of 8x8x4
Where data follows a Multinomial Normal Distribution with parameter \( \rho \) \( \mathcal{N}(0, \Sigma_\rho) \)
Add coherence between voxels

3 values for \( \rho \) [0.1, 0.5, 0.9]
Bigger \( \rho \) = stronger coherence

100 data samples (subjects) in the matrix \( \mathbf{A} \)

\[ \mathbf{x}: \]
Same structure as \( \mathbf{a}_i \)
Each block has a different strength [0.1 0.2 0.3 0.4]
Will determine how much the signal in this region contributes to the data label \( y_i \)

\[ \mathbf{y}: \]
Calculated using the value of the logistic regression distribution \( p(y|\mathbf{a}_i) \)
+1/-1 labels using the Bernoulli distribution \( \mathcal{B}(1,p) \)

\[
p(y_i|\mathbf{a}_i, \mathbf{x}, b) = \frac{1}{1 + \exp(-y_i(\mathbf{a}_i \mathbf{x} + b))}
\]
• **Accuracy of regions:** sparse classification is best in determining stable and accurate regions of difference

• **Classification accuracy:** Eigenanatomy slightly better results
Real MRI data

MS study: Investigating the role of iron and atrophy in MS

37 RRMS patients (6 males) and 37 matched controls
Age: RRMS 35.63 (std 9.2)
   Controls 35.69 (std 9.0) pval .97

MRI data at 4.7T: T1w, R2*, QSM (284 x 222 x 84 at .9 x .9 x 2 mm)
All data is normalized to an in-house unbiased template (nonlinear registration on T1w and QSM using ANTs)

Target regions: 4 subcortical deep GM structures
Only points inside this mask are considered in all methods

“Iron”: R2* data
Atrophy: logDetJac with respect to template
Note on data processing

High field MRI at 4.7T
• Several imaging modalities
  low contrast T1w

> standard segmentation methods are suboptimal
  FSL
Note on data processing

High field MRI at 4.7T
- Several imaging modalities
  low contrast T1w

- Quantitative iron sensitive MRI:
  $R^2*$ mapping  QSM

> standard segmentation methods are suboptimal
  FSL
Note on data processing

High field MRI at 4.7T
- Several imaging modalities
  low contrast T1w

Quantitative iron sensitive MRI:
- R2* mapping
- QSM

> standard segmentation methods are suboptimal
  FSL

Our own pipeline
Multi-atlas segmentation

Heckemann et al. Neurim 2005

- 10 manually segmented controls
- Nonlinear registration based on T1w+R2*+QSM (SyN ANTS)
- Label fusion
Multi-atlas segmentation

[Heckemann et al Neurio 2005][...]

- 10 manually segmented controls
- Nonlinear registration based on T1w+R2*+QSM (SyN ANTS)
- Label fusion

Nonlinear registration
Label fusion

MALF
FSL
Unbiased atlas

Use the same 10 controls to create an unbiased template
Iterative method from [Guimond et al. CVIU 2000]
Unbiased atlas

Use the same 10 controls to create an unbiased template. Iterative method from [Guimond et al. CVIU 2000]

All data is normalized into the space of the atlas by nonlinear registration (SyN) based on T1w, QSM and the MALF segmentation.

Look for regions that differentiate MS patients cs Control based on Iron (R2*) measurements and Atrophy (LogDetJac of deformations)
Significance of detected regions

Classification accuracy

Stability of regions

Regions: sparse classification detects most stable, accurate and significant regions.

Classification accuracy:
Also best for real data.
Extensions - data

**Vector data**
- Multi modalities images

**Tensor images**
(log > vectors space)

**Shape data**
as 3D points
Extensions - data

**Vector data**
- Multi modalities images

**Tensor images**
- (logEuclidean > vector space)

**Shape data**
- as 3D points

$\mathbf{A}_x \times \mathbf{A}_y \times \mathbf{A}_z$
- three sets of coefficients
- same entries should be zero - **group sparsity**
Extensions - regularization

Vector data

$$\min_x \|Ax - y\|^2_2 + \lambda_1 \|x\|_1 + \lambda_2 \|\nabla x\|^2_2$$

Discretized on the surface mesh

More general graph-based constraints for the image-regularization?

ex. diffusion tensors give regularization along brain fibres
Extensions – formulation

- Other type of discriminative energies: ex SVMs
- Deep learning? Convolutional nets?

ex. [Brosh et al. MICCAI 2014] Deep belief network used for generative learning of brain atrophy manifold
> how do we impose image-based regularization (compactness of features) Is convolutional enough?
MS data – big picture

Gray matter:

Atrophy defined on shapes

“iron” as voxel-based functional data
MS data – big picture

Gray matter:
- Atrophy defined on shapes
- "iron" as voxel-based functional data

White matter:
- Lesions
- Degradation of fibres (DTI data)
Gray matter:

Atrophy defined on shapes

“iron” as voxel-based functional data

White matter:

lesions

Degradation of fibres (DTI data)

How are all these parallel processes interacting?
How are they related to disease (group study, relate to disease duration, disease severity)?

All this data is nonlinearly related to aging
MS data – big picture

Gray matter:
- Atrophy defined on shapes
- “iron” as voxel-based functional data

White matter:
- Lesions
- Degradation of fibres (DTI data)

How are all these parallel processes interacting?
How are they related to disease (group study, relate to disease duration, disease severity)?

All this data is nonlinearly related to aging

THANK YOU