High Dimensional Data Analysis with Applications in IMS and fMRI Processing

Don Hong
Department of Mathematical Sciences
Center for Computational Sciences
Middle Tennessee State University
Murfreesboro, Tennessee, USA
http://capone.mtsu.edu/dhong
Outline

- Hyper-spectral imaging type data: IMS and fMRI
- Challenges in HSI type data processing
- Software package for IMS: IMSmining
- Markov random fields for IMS and Fusion IMS & Microscopy
- GLM, probabilistic group ICA and Multi-task learning for fMRI data analysis
HSI Data

- Hyperspectral images: images produced by imaging spectrometers; unlike conventional color cameras that only capture light in just RGB three spectral windows, hyperspectral cameras can capture an entire section of the electromagnetic spectrum at every pixel and collect high resolution spectral detail over a large spatial and broad wavelength region.
Hyperspectral Cube

Hyperspectral Imaging (HSI) Type: IMS and fMRI

- Data Cubes
- Spectral
- Spatial
- Medical/biological specifics
Example-1: Imaging Mass Spectrometry (IMS) Data

Plas R. V. et al. 2007. Imaging Mass Spectrometry Based Exploration of Biochemical Tissue Composition using Peak Intensity Weighted PCA.
MALDI-TOF

Mass spectra for each x,y coordinate

Single m/z value

MALDI-TOF MS technique was awarded the 2002 Nobel Prize in Chemistry
IMS Proteomic Cancer Data

Zhang and Hong, Statistics in Medicine, 30 (2011), 753-768
Example-2: MRI \xrightarrow{fMRI}

- **Magnetic resonance imaging (MRI)**, different from **computed tomography (CT)** or **X-rays**, can visualize internal structures of the body in detail, provide good **contrast** between the different **soft tissues** of the body, and thus, become very useful in imaging the **brain**, **muscles**, the **heart**, and **cancers**.

- Reflecting the fundamental importance and applicability of MRI in medicine, **Paul Lauterbur** of the **University of Illinois at Urbana-Champaign** and **Sir Peter Mansfield** of the **University of Nottingham** were **awarded the 2003 Nobel Prize in Medicine** for their "discoveries concerning magnetic resonance imaging".
fMRI

- **Functional magnetic resonance imaging** or **functional MRI** (fMRI) is an MRI procedure that measures brain activity by detecting associated changes in blood flow.

- The main principle of fMRI is using the difference of blood oxygenation level dependent (BOLD) to observe the local changes of deoxyhemoglobin concentration in the brain vasculature.

- Then, fMRI data can be measured at voxel by voxel, with a voxel representing a three dimensional cube unit inside the whole brain.
The fMRI Data Cube is of HSI Type

[http://bjr.birjournals.org/content/77/suppl_2/S167/F1.large.jpg]
Important Characteristics and Challenges in HSI Type Data Processing

- Data sets are very high dimensional
- Many challenges in data processing and data analysis: data preprocessing, dimension reduction, spatial information incorporation for feature selection and pattern recognition
- Requiring interdisciplinary collaboration
- Advanced mathematical tools and statistical techniques can not only provide significance analysis of experimental data sets but also can help in finding new data features/patterns, guiding biological experiments designs, as well as leading computational tools development.
Important Characteristics and Challenges in HSI Type Medical Data Processing

- Different from typical image processing emphasis, in HSI type medical data processing, we concentrate on the following 4Cs
  - Content of biology
  - Contrast between groups such as normal data and cancer data
  - Characteristic/feature/biomarker discovery
  - Classification accuracy improvement
I. MS/IMS Data Processing

- MS Data Preprocessing
  *with Shuo Chen and Yu Shyr*

- IMS Data Biomarker Discovery and Classification
  *with FQ Zhang and Lu Xiong*

- IMSmining Software Development
  *with JS Liang, FQ Zhang, and JC Zou*
IMSmining Software Package

IMSmining is a newly developed software package by our research group for IMS data processing using statistical methods. It contains the following functions:

1) Data visualization
   User can view the spectrum of a single pixel, the average spectrum of an area and intensity distribution matrix at a fixed m/z value.

2) Biomarkers selection and classification.
   Statistical algorithm selections include PCA+SVM/LDA, EN/WEN, SPCA+SVM/LDA.

   The graphical user interface is very friendly and convenient. Users can directly select an area of pixels from the ion image as the training data sets, choose the data sets for testing or classification from folders, and export data sets, mean spectrum, selected m/z list for later usage.

   Another great convenience is the mutual response between the ion image and the spectrum figure.
IMSmining

Work flow:
1) Import the training data and select one algorithm.
2) Use mouse to drag square areas of cancer and noncancer.
3) Import the testing data.

Results:
1) Biomarker.
2) Visualize testing data and classification results.
3) Accuracy and Sensitivity.
Import the data from .mat file or .txt folder. Export the biomarker.
Execute the algorithms including PCA, SVM, EN, WEN, SPCA.
View the spectrum of a single pixel or the average spectrum of an area.
(Toolbar) Save, Zoom in, Zoom out, Drag, and Rotate.
IMSmining

View and intensity distribution matrix at a fixed m/z value.
IMSmining

View the spectrum at a fixed pixel.
User also can click the mouse to view the intensity distribution matrix or spectrum directly.
EN & WEN

Test Data

EN
SPCA+SVM/LDA

However, by using PCA, each principal component is a linear combination of all the original variables. SPCA uses a regression-style of PCA and adds a lasso term to produce sparse coefficients.

We also include SPCA for IMS processing.
## Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Time(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA+SVM</td>
<td>84.88%</td>
<td>88.23%</td>
<td>16.49</td>
</tr>
<tr>
<td>EN/WEN</td>
<td>94.02%</td>
<td>90.04%</td>
<td>29.56</td>
</tr>
<tr>
<td>SPCA+SVM</td>
<td>95.43%</td>
<td>93.75%</td>
<td>14.67</td>
</tr>
</tbody>
</table>

- PCA+SVM/LDA is the worst one.
- EN and WEN perform better and are almost the same based on this dataset.
- SPCA+SVM/LDA is the best one considering results and time consuming.
Help in Discovery of New Biomarkers:

EN4IMS/WEN models found more important biomarkers compared to PCA method. The peaks (m/z 6700, 8380, 10952, 14788) known as cancer biomarkers are on EN4IMS/WEN list but not on PCA list.

Spatial intensity distribution of 4 selected biomarkers. The x- and y- dimensions are the spatial dimensions and intensities of the selected m/z values are represented by the color.
IMSmining also includes MRA for IMS Data Biomarker Selection and Classification

- **Basic Idea and Procedure:**
  1. Apply wavelet transform to IMS data
  2. The idea of wavelet pyramid method for image matching was then applied for biomarker selection, in which Jaccard similarity was used to measure the similarity of wavelet coefficients and determine feature vectors to serve as the biomarkers.
  3. Last, the Naive Bayes classifier was used for classification based on feature vectors in terms of wavelet coefficients.
Biomarker Selection using Wavelets

Statistically compare using Jaccard similarity

Cancer wavelet coefficients group

Non-cancer wavelet coefficients group

resolution, J

position K
Spatial Information Consideration

The state (cancer or non-cancer) of a pixel is highly determined by the configuration of its neighboring system.

To fully utilize IMS data, it is desirable to not only identify the peaks of the spectrum within individual pixels but also to study correlation and distribution using the spatial information for the entire image cube.
Markov Random Fields

Local property (the Markovianity)

Define

\[ S - \text{set of lattice points} \]
\[ s - \text{a lattice point, } s \in S \]
\[ X_s - \text{the value of } X \text{ at } s \]
\[ \partial s - \text{the neighboring points of } s \]

\[ p(x_s \mid x_r \text{ for } r \neq s) = p(x_s \mid x_{\partial r}) \]
MCMC-MRF Algorithm for IMS Data Processing

MS Data Preprocessing Papers


IMS Cancer Data Processing
Papers

- **EN4IMS**
  

- **WEN**
  

- **MRA Based method**
  
  
Experiment Data

Training data

Test data
Image fusion of mass spectrometry and microscopy

doi:10.1038/nmeth.3296
Fusion IMS & Microscopy

II. fMRI Data Processing

- The aims of fMRI analysis include: localizing regions of the brain activated by a certain task; determining distributed networks that correspond to brain function; and making predictions about psychological or disease states.

- Also, modeling brain connectivity, visualizing and localizing activation, and analyzing region of interest.
Statistical Analysis of fMRI Data
fMRI Data Processing

- Two popular models to process the fMRI data are General Linear Model (GLM) and Independent Component Analysis (ICA). Both methods need to incorporate adaptivity for fMRI study.
- The fMRI design that the BOLD response behaved as a linear time invariant system is crucial for simplifying the analysis by allowing the using of GLM.
- Since 2000, new approaches that analyzes patterns of activity rather than the response at voxels are developed including using pattern information analysis and machine learning.
General Linear Model

1. Model (1 or more Regressors)

2. Data

3. Fitting the Model to the Data at each voxel

\[ y(i) = \hat{\beta}_0 + \sum_{j=1}^{M} \hat{\beta}_j x_j(i) + e(i) \]
RESOURCES

- Context determines the meaning and interpretation of the word “network” in brain imaging analysis.

- GLM and seed-based methods define a network as a subset of voxels whose time series are significantly correlated with a reference signal.

- ICA defines a network as a subset of voxels whose time series are significantly correlated with the estimated ICA time course.
GLM usually needs to know the design patterns of the experiments. It is simple to extend the application of GLM from single subject analysis to the group analysis. However, for some specific design experiment, we merely know the patterns.

We can use ICA to separate the independent patterns and the spatial maps from the BOLD signals. Temporal concatenation is the most popular way to organize the group data.

We can obtain the group specific temporal responses and common spatial maps via the tensor model.
PCA and ICA

PCA finds directions of maximal variance

ICA finds directions which maximize independence

ICA for fMRI

- Let \( i = 1, \ldots, N \) represent \( N \) subjects, \( t = 1, \ldots, T \) the \( T \) time points, and \( v = 1, \ldots, V \) the \( V \) voxels. Use \( X \) to represent the temporal concatenation of fMRI data of \( N \) subjects (groups). The group ICA model can be expressed as:

\[
X = MS + E,
\]

- where each column of \( S \) is a \( q \times 1 \) vector containing \( q \) statistically independent non-Gaussian spatial source signals at the voxel \( v \), \( M \) is a \( TN \times q \) matrix that contains \( q \) independent spatial source signals of the observed multi-subject fMRI images, each column of \( E \) is a \( TN \times 1 \) vector representing the \( N \) subjects’ noise at voxel \( v \) following a multivariate normal distribution \( \text{MVN}(0, IN \otimes \Sigma_v) \) where \( \Sigma_v \) is a \( T \times T \) error covariance.
Figure: ICA explanation.
To be better to investigate differentiations between groups of subjects in terms of a common time courses with subject loading coefficients under the ICA spatial map, following a similar procedure of Probabilistic PCA, JS Liang at MTSU COMS PhD program proposed a so-called multi-group Probabilistic tensorial ICA algorithm for fMRI data analysis to improve computing efficiency and accuracy of the group ICA algorithm.
Multi-Group PTICA for fMRI

- Algorithm Outline:

1. Modify the original data $X$ time course to be normalized to zero mean and unit variance.

2. Reduce the dimension of the data at subject level using PCA. Use the appropriate matrix $H$, we have

$$HX = HMS + HE.$$ 

This can be deemed as a new equation of the same format

$$X = MS + E.$$ 

3. Use Laplace approximation or minimum description length (MDL) to estimate the number of the latent independent components to extract reasonable features from the mixed data in the following steps.
Algorithm Outline (cont.):

4. Use PICA approach for the decomposition of data $X$ into a mixing matrix $M$ and spatial maps $S$.

$$M_{ML} = U_q(\Lambda_q - \sigma^2 I_q)^{1/2}Q^t$$

$$S_{ML} = (A^t A)^{-1} A^t X$$

$$\sigma^2_{ML} = \frac{1}{(NT - q)} \sum_{l=q+1}^{NT} \lambda_l$$

where $X = U(N\Lambda)^{1/2}V$. $U_q$ and $\Lambda_q$ contain the first $q$ eigenvectors and eigenvalues of $U$ and $\Lambda$, and $Q$ denotes a $q \times q$ orthogonal rotation matrix.

5. For each group, use rank-1 approximation via SVD to decompose the mixing matrix $M$ to $C \otimes A$, where $\otimes$ is Khatri-Rao Product. Therefore, for each group, we try to find the common time course $A$ and the corresponding subject loading $C$. 
6. Iterate the above steps until convergence.

7. Use a reconstruction method to rebuild the time courses $A$ and spatial maps $S$.

8. Apply some inference methods like modified $z$ maps to detect the activation maps.

9. Compare the time courses between the groups to find the potential ICs.

Suppose we have two groups: the contrast group and the group with Autism. Two groups have two time courses of the same independent spatial map. Calculate the correlate coefficient $r$ between them. If $r$ is quite small, we intend to believe this IC is associated with the illness.
Applications

- Autism data analysis: not only compare patient group with normal group, the corresponding model can be used to study and compare groups based on severity scores to seek possible new biomarkers

- Alzheimer Disease data study
Ph.D. student, Xin Yang (co-supervising with Qiang Wu), applies multitask learning scheme for fMRI data analysis.

The idea is to find a better pattern representation by using multitask learning technique defined on voxels.

In the subsequent slides, index $t$ is for the task labels.
Spatial Regularization Multi-task Learning

• In MTL regression, there are $T \geq 2$ tasks. Assume the $t$-th task has the data matrix $X_t$ and response vector $Y_t$ which are linked by

\[ Y_t = X_t \beta_t + E_t \]

• To code the spatial information, we first define a neighborhood system. It is defined by user and may be quite data dependent.

![4 Neighborhood](image1.png)  ![8 Neighborhood](image2.png)

(a) 4 Neighborhood  (b) 8 Neighborhood

**Figure**: Neighborhood Structure for each task
Spatial Regularization Multi-task Learning

- Based on the neighborhood system, we define task similarity coefficient by

\[ w_{tk} = \begin{cases} 
1, & \text{if task } k \text{ is a neighborhood of task } t. \\
0, & \text{if task } k \text{ is not a neighborhood of task } t. 
\end{cases} \]

- We assume the neighborhood system is symmetrically defined so that \( w_{tk} = w_{kt} \). The penalty term for spatial regularization is defined by

\[ \lambda_s \sum_{t,k=1}^{T} w_{tk} \| \beta_t - \beta_k \|_2^2 \]

When \( \lambda_s \) becomes large, it forces the neighboring tasks are very close while as \( \lambda_s \) tends to 0, the tasks are treated as independent.
Spatial Regularization MTL: Spatial Ridge

- When we learn all $T$ ridge regression problem simultaneously and apply the spatial penalty, the resulted MTL learning algorithm, called spatial ridge regression algorithm, takes the form.

$$
\hat{B}_{SR} = \arg \min_{B} \left\{ \sum_{t=1}^{T} \left\| Y_t - X_t\beta_t \right\|_2^2 + \lambda_2 \sum_{t=1}^{T} \left\| \beta_t \right\|_2^2 + \lambda_s \sum_{t,k=1}^{T} \omega_{tk} \left\| \beta_t - \beta_k \right\|_2^2 \right\} \tag{9}
$$
Multi-task Learning for fMRI

- From the simulation and real data results, we see that MTL algorithms achieve much better results than STL algorithms.
- Regularized MTL algorithm considers all task’s average instead of tasks with similar features.
Acknowledgement

- Mass Spectrometry Research Center at Vanderbilt University School of Medicine.
- Institute of Imaging Science, Vanderbilt University
- Department of Child & Adolescent Psychiatry, Vanderbilt University
- Institute of Images and Pattern Recognition, North China University of Technology
The Journal of Health & Medical Informatics (JHMI) is a scholarly journal that integrates the information and communication related to healthcare, medicine and their applications in medical information systems.

Don Hong, Editor-in-Chief,
Middle Tennessee State University, Murfreesboro, TN

David Randall, Co-Editor-in-Chief,
the Consumer Driven Health Care Institute, Washington DC, USA

Submit manuscript at http://editorialmanager.com/omicsgroup/ or send as an e-mail attachment to the Editorial Office at editor.jhmi@omicsonline.org
Thank you !!!