Big Data Integration in Biomedical Studies: A Few Personal Views and Experiences

Hongtu Zhu, Ph.D
Department of Biostatistics‡ and Biomedical Research Imaging Center‡
The University of North Carolina at Chapel Hill,
Chapel Hill, NC 27599, USA
Outline

• Big Data
• BIAS and Big Data Integration
• Image-on-Scalar Models
• Image-on-Genetic Association Models
• Prediction Models
What is ‘Big Data’?

$5V = \text{Volume, Velocity, Variety, Value, and Veracity}$

The size of big data is beyond the ability of commonly used software tools to capture, manage, and process within a tolerable elapsed time.

- Alzheimer’s Disease Neuroimaging Initiative (US$134 millions)
- Philadelphia Neurodevelopmental Cohort (PNC)
- Human Connectome Project (HCP)

Big Data in Boxes
How to promote statistics in ‘Big Data’ industry?

• Closely collaborate with people who are collecting ‘Big Data’

• Work as a team to develop new methods, packages, and textbooks with nice case studies

• Organize more workshops and short courses

• Train next-generation statisticians: training grants and new courses a data scientist; an excellent programmer; an applied mathematician
How to make important contributions to ‘Big Data’?

• Start with a few big data bases

• Start with a few methodological and clinical projects

• Develop a package with a set of good computational and statistical tools to efficiently extract important information from large Big data
BIAS: Biostatistics and Imaging Analysis and Big Data Integration
BIAS: Biostatistics and Imaging Analysis Lab

Man Power
Computer Power
Programming Power
Statistics
Mathematics

http://www.bios.unc.edu/research/bias/
The Brain Research through Advancing Innovative Neurotechnologies or BRAIN, aims to reconstruct the activity of every single neuron as they fire simultaneously in different brain circuits, or perhaps even whole brains.
Big Neuroimaging Data

NIH normal brain development
1000 Functional Connectome Project
Alzheimer’s Disease Neuroimaging Initiative
National Database for Autism Research (NDAR)
Human Connectome Project
Philadelphia Neurodevelopmental Cohort
Genome superstruct Project

www.guysandstthomas.nhs.uk/.../T/Twins400.jpg
Complex Study Design

cross-sectional studies; clustered studies including longitudinal and twin/familial studies;

4D Longitudinal Data

\[ \rho_{1 \rightarrow 3} \quad \rho_{3 \rightarrow 5} \]
Neuroimaging Applications

**Structural MRI**
- Variety of acquisitions
- Measurement basics
- Limitations & artefacts
- Analysis principles
- Acquisition tips

**Diffusion MRI**

**Functional MRI (task)**

**Functional MRI (resting)**
Complex Data Structure

Multivariate Imaging Measures
Smooth Functional Imaging Measures
Whole-brain Imaging Measures
4D-Time Series Imaging Measures
Big Data Integration

Medical Informatics & Management

Disease
   Etiology
   Prevention
   Treatment

Medical Industry
   Care
   Policy
   System
   Science
   Insurance
   Economics
   Pharmaceutical
Big Data Integration

E: environmental factors
G: genetic markers
D: disease

Selection

http://en.wikipedia.org/wiki/DNA_sequence
Image-on-Scalar Models
E: environmental factors

G: genetic markers

D: disease

Selection

http://en.wikipedia.org/wiki/DNA_sequence
Strategic Objective 1: Promote Discovery in the Brain and Behavioral Sciences to Fuel Research on the Causes of Mental Disorders

Identifying and validating high sensitivity and specificity biomarkers that define valid subtypes of the major mental illnesses.

Strategic Objective 2: Chart Mental Illness Trajectories to Determine When, Where, and How to Intervene

Conducting longitudinal studies that track changes in behavior with brain structure, connectivity, and function, in order to characterize the progression from primary changes to subsequent clinical presentation, and to identify predictors of divergence from the typical trajectory.
Smoothed Functional Data

Covariates (e.g., age, gender, diagnostic)
Case 1: DTI Fiber Tract Data

- Diffusion properties (e.g., FA, RA)
  \[ Y_i(s_j) = (y_{i,1}(s_j), \cdots, y_{i,m}(s_j))^T \]
- Grids \[ \{ s_1, \cdots, s_{n_G} \} \]
- Covariates (e.g., age, gender, diagnostic)
  \[ x_1, \cdots, x_n \]
Objectives:
Dynamic functional effects of covariates of interest on functional response.
Ex 1: Longitudinal Tract Data

DTImaging parameters:

- TR/TE = 5200/73 ms
- Slice thickness = 2mm
- In-plane resolution = 2x2 mm^2
- b = 1000 s/mm^2
- One reference scan b = 0 s/mm^2
- Repeated 5 times when 6 gradient directions applied.
Ex 1: Longitudinal Tract Data

(a) Arclength

(b) Arclength

(c) Arclength

(a) Age (days)

(b) Age (days)

(c) Age (days)
Ex 1: Longitudinal Tract Data
Neuroimaging Data with Discontinuity

Noisy Piecewise Smooth Function with Unknown Jumps and Edges

Covariates (e.g., age, gender, diagnostic, stimulus)
Noisy Piecewise Smooth Functions with Unknown Jumps and Edges

**Image** is the point or set of points in the range corresponding to a designated point in the domain of a given function.

\[ \Omega \text{ is a compact set. } \tilde{x} \in \Omega \subseteq R^k \]

\[ f(\tilde{x}) \in M \subseteq R^m \]

\[ f : \Omega \rightarrow M \subseteq R^m \]

\[ \int_{\Omega} \|f(\tilde{x})\|^k \, d\tilde{x} < \infty \text{ for some } k > 0 \]
M2: Spatial Varying Coefficient Model

Decomposition:

\[ y_i(d) = f(x_i, B(d) + \eta_i(d)) + \varepsilon_i(d), d \in D \]

- **Piecewise Smooth Varying Coefficients**
  \[ B(d) \in L^K \]
- **Long-range Correlation**
  \[ \eta_{ij}(\bullet) \sim SP(0, \Sigma_\eta) \]
- **Short-range Correlation**
  \[ \varepsilon_{ij}(\bullet) \sim SP(0, \Sigma_\varepsilon) \]

**Covariance operator:**

\[ \Sigma_y(d, d') = \Sigma_\eta(d, d') + \Sigma_\varepsilon(d, d) \]

Li, Zhu, Shen, Lin, Gilmore, and Ibrahim (2011). JRSSB.
Zhu, Fan, and Kong (2014) JASA

The UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
Spatial Varying Coefficient Model

Cartoon Model

- Disjoint Partition:
  \[ D = \bigcup_{l=1}^{L} D_l \quad \text{and} \quad D_l \cap D_{l'} = \emptyset \]

- Piecewise Smoothness: Lipschitz condition

- Smoothed Boundary

- Local Patch

- Degree of Jumps

\[ B_k (d) \]
Kernel-based Smoothing Methods

\[ y = f + \varepsilon; \quad \varepsilon \text{ uncorrelated, mean}=0, \text{var}=\sigma^2 \]

Estimate \( f_i \) as a weighted average of the noisy pixels:

\[ \hat{f}_i = \sum_j w_{i,j} y_j \]

Arias-Casto, Salmon, Willett (2011)
- Local constant/linear
- Yaroslavsky/Bilateral Filter
- Nonlocal Means
- PS
Kernel-based Smoothing Methods

Propogation-Seperation Method

Features

- Increasing Bandwidth
\[ 0 < h_0 < h_1 < \cdots < h_S = r_0 \]
- Adaptive Weights
- Adaptive Estimates
Simulation

True Image

SVCM

Initial Estimate in SVCM

Estimate with LF and r=0

Estimate with LF and r=1

Estimate with LF and r=2
Simulation

Bias with SVCM

Bias with LF and r=0

MSE with SVCM

MSE with LF and r=0

Bias with LF and r=1

Bias with LF and r=2

MSE with LF and r=1

MSE with LF and r=2
EX2: ADNI PET Data

- Data were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database.
- We consider PET scans obtained at baseline, 6 months, and 12 months.
- Subjects are classified as having mild cognitive impairment (MCI), as AD patients, or as Normal Controls (NC).

<table>
<thead>
<tr>
<th>Diagnostic status</th>
<th>age (years)</th>
<th>N male</th>
<th>N female</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>75.9± 6.9</td>
<td>34</td>
<td>17</td>
</tr>
<tr>
<td>MCI</td>
<td>76.3± 7.3</td>
<td>33</td>
<td>25</td>
</tr>
<tr>
<td>NC</td>
<td>77.0± 4.2</td>
<td>30</td>
<td>20</td>
</tr>
</tbody>
</table>

- We randomly chose 80 subjects for the training set to develop the prediction model.
- We predicted the PET scans at month 12, based on the baseline and 6-month scans for 79 subjects in the test set.
- We used gender, diagnostic status (MCI, AD, NC), and age (55-90 years) as covariates for the semi-parametric model.

SGPP. NeuroImage
EX2: ADNI PET Data

Figure: Observed (upper panel) and predicted (bottom panel) PET images at month 12 for (a) an AD patient, (b) an MCI subject, and (c) a NC subject. One selected slice is shown.
EX2: ADNI PET Data

Table: rtMSPE for ADNI PET images

<table>
<thead>
<tr>
<th>Model</th>
<th>rtMSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semi-parametric model</td>
<td>0.0692</td>
</tr>
<tr>
<td>Semi-parametric model + FPCA</td>
<td>0.0550</td>
</tr>
<tr>
<td>Semi-parametric model + FPCA + Spatial-temporal model</td>
<td>0.0354</td>
</tr>
</tbody>
</table>

Figure: rtMSPE maps for prediction of ADNI PET images at month 12 for 79 test subjects. Selected slices are shown for (a) Semi-parametric model; (b) Semi-parametric model + FPCA; (c) Semi-parametric model + FPCA + Spatial-temporal model.
Image-on-Genetic Association Models
The NIMH Strategic Plan

Strategic Objective 1: Promote Discovery in the Brain and Behavioral Sciences to Fuel Research on the Causes of Mental Disorders

Identify the genetic and environmental factors associated with mental disorders.

Strategic Objective 2: Chart Mental Illness Trajectories to Determine When, Where, and How to Intervene

When identifying behavioral, neural, and/or genetic markers along the trajectory of illness, design the studies to consider variation in relation to age, sex, gender, race, ethnicity, and other important socio-demographic factors.
Big Data Integration

E: environmental factors

G: genetic markers

D: disease

Selection

http://en.wikipedia.org/wiki/DNA_sequence
<table>
<thead>
<tr>
<th>Imaging</th>
<th>Candidate ROI</th>
<th>Many ROI</th>
<th>Voxelwise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetics</td>
<td><img src="image1.png" alt="image" /></td>
<td><img src="image2.png" alt="image" /></td>
<td><img src="image3.png" alt="image" /></td>
</tr>
<tr>
<td>Candidate SNP</td>
<td><img src="image4.png" alt="image" /></td>
<td><img src="image5.png" alt="image" /></td>
<td><img src="image6.png" alt="image" /></td>
</tr>
<tr>
<td>Candidate Gene</td>
<td><img src="image7.png" alt="image" /></td>
<td><img src="image8.png" alt="image" /></td>
<td><img src="image9.png" alt="image" /></td>
</tr>
<tr>
<td>Genome-wide SNP</td>
<td><img src="image10.png" alt="image" /></td>
<td><img src="image11.png" alt="image" /></td>
<td><img src="image12.png" alt="image" /></td>
</tr>
<tr>
<td>Genome-wide Gene</td>
<td><img src="image13.png" alt="image" /></td>
<td><img src="image14.png" alt="image" /></td>
<td><img src="image15.png" alt="image" /></td>
</tr>
</tbody>
</table>

Hibar, et al. HBM 2012
Data Structure

Imaging:

Genetic:

The UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
M3: High Dimensional Regression Model

Data: \(\{(Y_i, X_i) : i = 1, \cdots, n\}\)

\(Y_i = \{y_i(v) : v \in V\}\) \(\{X_i(g) : g \in G_0\}\)

\[
\begin{align*}
Y & \quad \text{Phenotype} \\
\begin{array}{c}
\text{Genotype} \\
\end{array} & \quad X \\
\end{align*}
\]

\[
\begin{array}{c}
\begin{array}{c}
\text{Genotype} \\
\end{array} & \quad B \\
\end{array} + E
\]

\[
\begin{array}{c}
\begin{array}{c}
\text{Genotype} \\
\end{array} & \quad n \times p_y \\
\end{array}
\]

\[
\begin{array}{c}
\begin{array}{c}
\text{Genotype} \\
\end{array} & \quad p_x \times p_y \\
\end{array}
\]

\[
\begin{array}{c}
\begin{array}{c}
\text{Genotype} \\
\end{array} & \quad (p_x, n, p_y)
\end{array}
\]
Sparse Projection Regression Model

Let \( W = [w_1, \cdots, w_k] \), then a projection regression model is given by:

\[
W^T y_i = (BW)^T x_i + W^T e_i = \beta_w^T x_i + \varepsilon_i
\]

Hypothesis problem reduces to:

\[
H_{0W} : C\beta_w = b_0 \quad v.s. \quad H_{1W} : C\beta_w \neq b_0
\]
where \( C\beta_w = CBW \) and \( b_0 = B_0W \)

How to determine an ‘optimal’ \( W \)?

Sun, Zhu, Liu, and Ibrahim (2014) JASA
Sparse Projection Regression Model

- We show that this is achieved by optimizing the following generalized heritability ratio (GHR):

\[
GHR(w; \mathbf{C}) = \frac{w^T (\tilde{B}_1 - B_0)^T \tilde{S}_{\tilde{X}_1} (\tilde{B}_1 - B_0) w}{w^T \Sigma_R w} = \frac{w^T \Sigma_C w}{w^T \Sigma_R w}
\]

- High Dimensional Setting
- noise accumulation
  - ill-conditioned sample covariance estimator: \( \hat{\Sigma}_R \)

- Sparse Projection Regression Model is proposed as following:

\[
\arg\max \left\{ \frac{w^T \hat{\Sigma}_C w}{w^T \hat{\Sigma}_R w} \right\} \quad \text{s.t.} \quad \|w\|_1 \leq t
\]
Sparse and Low-rank Representation

Sparsity and Structure on B.

\[ B = b_X \times b_Y \]

Low Rank

\[ p_\lambda(B) = p_\lambda(b_X) + p_\lambda(b_Y) + p_\lambda(E_B) \]

Sparsity

Regularization Methods

- Lasso 1, 2, 3, …
- SCAD, MCP, …

\[ \hat{\theta} \in \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda_n \sum_{j=1}^{p} |\theta_j| \]

Shen, Shen, and Zhu (201?)

The UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
Factor Model

\[ E_{i}^{p_y \times 1} = \Lambda^{p_y \times q} \xi_{i}^{q \times 1} + \eta_{i}^{p_y \times 1} \]

\[ \Sigma_{E} = \Lambda^{T} \Sigma_{\eta} \Lambda \]

Zhu, Zakaria, Lu, and Ibrahim (2014) JASA
M4: Voxel-wise GWAS

~1000 subjects

~30,000 voxels in the brain

~600,000 genetic markers (SNPs)

1.8 x 10^{10} tests!

Hibar, et al. HBM 2012
M4: Voxel-wise GWAS

Fast Sure-Independence Screening Procedure

Huang, .... and Zhu (2014)
EX4: Whole Brain-GWAS
Prediction Models
Alzheimers Disease Big Data DREAM Challenge 1

Its goal is to apply an open science approach to rapidly identify accurate predictive AD biomarkers that can be used by the scientific, industrial and regulatory communities to improve AD diagnosis and treatment.

**Sub 1:** Predict the change in cognitive scores 24 months after initial assessment.

**Sub 2:** Predict the set of cognitively normal individuals whose biomarkers are suggestive of amyloid perturbation.

**Sub 3:** Classify individuals into diagnostic groups using MR imaging.
Big Data Integration

E: environmental factors

G: genetic markers

D: disease

http://en.wikipedia.org/wiki/DNA_sequence
HRM versus FRM

Data

\{(y_i, X_i): i = 1, \cdots, n\}

\[ X_i = \{X_i(d): d \in D\} \]

\[ y_i = \langle X_i, \theta \rangle + \varepsilon_i \]

Strategy 1: Discrete Approach
(High-dimension Regression Model (HRM))

\[ y = X \theta^* + w \]

Strategy 2: Functional Regression Model (FRM)

\[ y_i = \theta_0 + \int_D \theta(d)X_i(d)m(d) + \varepsilon_i \]
Key Conditions:

- Sparsity of $S$
- Restricted null-space property for design matrix $X$

\[
\hat{\theta} \in \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda_n \sum_{j=1}^{p} |\theta_j| \]
High-dimension Regression Model

Tensor Structure:
- Ultra-high dimensionality ($256^3$)
- Spatial structure

$
\begin{align*}
\mathbf{y} & \sim \mathbf{X} \mathbf{\theta}^* + \mathbf{w} \\
\mathbf{n} & \times \mathbf{p} \\
\text{Zhou, Li, and Zhu (2013)} & \\
\text{Li, Zhou, and Li (2013)} &
\end{align*}
$

CP decomposition

Tucker decomposition
Scalar-on-Image Models

Key Conditions:
- Tensor Approximation B
- Restricted space property for X and B
Scalar-on-Image Models

Strategy 2: Functional Approach

\[ y_i = \theta_0 + \int_{D} \theta(d)X_i(d)m(d) + \epsilon_i \]

\[ \theta(d) = \sum_{k=1}^{\infty} \theta_k \psi_k(d) \]

\[ y_i = \theta_0 + \sum_{k=1}^{\infty} \theta_k \int_{D} \psi_k(d)X_i(d)m(d) + \epsilon_i \]

Basis Methods: fixed and data-driven basis functions

\[ K_\theta = \{ \theta(d) = \sum_{k=1}^{\infty} \theta_k \psi_k(d) : (\theta_1, \cdots) \in \ell^2 \} \]

\[ C(d,d') = \text{Cov}(X(d),X(d')) = \sum_{k=1}^{\infty} \lambda_k \xi_k(d)\xi_k(d') \]
Key Conditions: an excellent set of basis functions

- **Sparsity of basis representation** \[ \{ \theta_k : k = 1, \ldots \} \]
- **Decay rate of spectral of** \( C \) \text{ or } \frac{1}{2} \frac{K^{1/2} CK^{1/2}}{C K^{1/2}}

\[
\theta(d) \approx \sum_{k=1}^{K} \theta_k \psi_k(d) \quad K \ll n
\]
Extensions

• M5: Functional linear Cox regression models
• M6: Generalized scalar-to-image regression models
• M7: Multiscale Functional Linear models
M5: Functional Linear Cox Regression Model

Data
\[ \{(y_i, X_i) : i = 1, \cdots, n\} \quad X_i = \{X_i(d) : d \in D\} \]
\[ y_i = \min(T_i, C_i) \quad T_i : \text{failure time}; \quad C_i : \text{censored time} \]

Model
\[ h(t) = f(t) / S(t) = h_0(t) \exp(z_i^T \gamma + \int_{s} X_i(s) \beta(s) ds) \]
\[ X_i(s) = \mu(s) + \sum_{j=1}^{\infty} \xi_{ij} \phi_j(s) + \varepsilon_i(s) \]

- Consistency
- Asymptotic distribution of score test
Mild Cognitive Impairment subjects

Interested in predicting the timing of an MCI patient that converts to AD by integrating the imaging data, the clinical variables, and genetic covariates.

Full Model: AUC=0.96
Partial Model: AUC=0.82
Data \( \{(y_i, X_i) : i = 1, \cdots, n\} \quad X_i = \{X_i(d) : d \in D\} \)

Model \( y \sim \) exponential family(\( \mu, \phi \))

\[ g(\mu) = \theta_0^T Z + < X, \beta_0 > \]

Estimation:

\[ \sum_{i=1}^{n} \ell(y_i; \mu(X_i; \gamma, \beta(\bullet))) + \lambda \| \beta \|_{TV} \]

Non-asymptotic Error Bound:

\[ \mathcal{R}_{2n} = \left\{ \mathbb{E}^* \left( \left\langle X^{(n+1)}, \hat{\beta} - \beta_0 \right\rangle \right)^2 \right\}^{1/2} , \]
M6: Generalized scalar-to-image regression models
M7: Multiscale Functional Linear models

Data \( \{(y_i, X_i) : i = 1, \ldots, n\} \quad X_i = \{X_i(d) : d \in D\} \)

Models

(A1) \( D = \bigcup_{k=1}^{K} D_k \bigcup D_0 \) \quad • Informative sets + Irrelevant set

(A2) \( y \perp \{X(d) : d \in D_0\} \)

(A3) \( y \sim p(\{X(d) : d \in D_1\}, \ldots, \{X(d) : d \in D_K\}) \)
Simulation I: Classification

Class 0

Class 1

\[ X_i(d) = \beta_0(d) + \beta_1(d)y_i + \varepsilon_i(d) \]

Type I

Type II

Type III

\( N(0,4) \)

Short-range correlation

Long-range correlation

0 White
1 Green
2 Red
Simulation I: Classification

Table 1: Misclassification rates for PCA and SWPCA under the different number of PCs.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Number of PCs</th>
<th>PCA</th>
<th>SWPCA1</th>
<th>SWPCA2</th>
<th>SWPCA3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>5</td>
<td>0.40</td>
<td>0.11</td>
<td>0.09</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.40</td>
<td>0.13</td>
<td>0.11</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.40</td>
<td>0.13</td>
<td>0.11</td>
<td>0.10</td>
</tr>
<tr>
<td>Type II</td>
<td>5</td>
<td>0.40</td>
<td>0.04</td>
<td>0.08</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.39</td>
<td>0.03</td>
<td>0.09</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.38</td>
<td>0.03</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td>Type III</td>
<td>5</td>
<td>0.40</td>
<td>0.13</td>
<td>0.10</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.41</td>
<td>0.13</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.41</td>
<td>0.13</td>
<td>0.10</td>
<td>0.10</td>
</tr>
</tbody>
</table>
Simulation I: Classification

<table>
<thead>
<tr>
<th>Noise</th>
<th>sLDA</th>
<th>sPLS</th>
<th>SLR</th>
<th>SVM</th>
<th>ROAD</th>
<th>PCA</th>
<th>SWPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>0.28</td>
<td>0.43</td>
<td>0.45</td>
<td>0.38</td>
<td>0.36</td>
<td>0.36</td>
<td>0.10</td>
</tr>
<tr>
<td>Type II</td>
<td>0.27</td>
<td>0.08</td>
<td>0.18</td>
<td>0.26</td>
<td>0.08</td>
<td>0.45</td>
<td>0.03</td>
</tr>
<tr>
<td>Type III</td>
<td>0.52</td>
<td>0.30</td>
<td>0.61</td>
<td>0.60</td>
<td>0.50</td>
<td>0.35</td>
<td>0.09</td>
</tr>
</tbody>
</table>

sLDA: sparse discriminant analysis
sPLS: sparse partial least squares analysis
SLR: sparse logistic regression
SVM: support vector machine
ROAD:

The UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
sification rates are respectively shown from the 2th to the 6th column of Table 2, which are larger than the classification errors by SWPCA. These simulation results furthermore confirm the better performance of SWPCA in classification.

1.2 Real Data Analysis

This section applies SWPCA to real data classification and further compares the performance of SWPCA with other classification methods. The real data is Alzheimer's Disease Neuroimaging Initiative (ADNI) pet data (Jagust et al., 2010). Alzheimer's disease (AD) is the most common form of dementia and results in the loss of memory, thinking and language skills. ADNI is a worldwide project, launched in 2003, and is to develop biological markers to track the progression of Alzheimer's disease (AD), to help the AD treatment. Many research groups, such as National Institutes of Health, the Food and Drug Administration (FDAs), and others, are collaborating on ADNI to achieve this goal.
The misclassification rates are respectively shown from the 2th to the 6th column of Table 2, which are larger than the classification errors by SWPCA in the last column. These simulation results furthermore confirm the better performance of SWPCA in classification.

1.2 Real Data Analysis

This section applies SWPCA into the classification analysis of real data and further compares the classification performance of SWPCA with other classification methods. The real data is Alzheimer's Disease Neuroimaging Initiative (ADNI) pet data (Jagust et al., 2010). Alzheimer's Disease Neuroimaging Initiative is a worldwide project, launched in 2003. The goal of ADNI project is to develop biological markers to track the progression of Alzheimer's disease (AD) and helps the AD treatment. Multiple research groups, including National Institutes of Health, the Food and Drug Administration, drug and medical-imaging companies, universities and nonprofit groups, have contributed their findings of the biological markers to the understanding of the progression of Alzheimer's disease in the human brain. Here we consider the real data, which contains 196 $79 \times 95 \times 69$ 3D-images.

Leave one out test: 196 $79 \times 95 \times 69$ 3D-images are split into a training set with 195 images and a test set of one image. Simulation is repeated 196 times, error is average error on predicting testing image.

Table 3: Results of Real Data: average misclassification rates.

<table>
<thead>
<tr>
<th>Method</th>
<th>Misclassification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>sLDA</td>
<td>0.255</td>
</tr>
<tr>
<td>sPLS</td>
<td>0.163</td>
</tr>
<tr>
<td>sLogistic</td>
<td>0.179</td>
</tr>
<tr>
<td>SVM</td>
<td>0.168</td>
</tr>
<tr>
<td>ROAD</td>
<td>0.189</td>
</tr>
<tr>
<td>PCA</td>
<td>0.194</td>
</tr>
<tr>
<td>SWPCA</td>
<td>0.117</td>
</tr>
</tbody>
</table>

94 AD subjects and 104 NC subjects

Table 3: Results of Real Data: average misclassification rates.
Thank You!!

ASA: Statistics in Imaging Section

SAMSI

2013 Neuroimaging Data Analysis
2015-2016 Challenges in Computational Neuroscience

July 27-31 Summer School    August 17-21 Opening Workshop

• Shape Analysis
• Spike Train Analysis
• Big Data Integration
• Compressed Sensing
• Functional Data Analysis

The UNIVERSITY of NORTH CAROLINA at CHAPEL HILL