Statistical Learning Methods for Neuroimaging Data Analysis with Applications

University of North Carolina at Chapel Hill

Hongtu Zhu

https://www.med.unc.edu/big-s2
Part I

A Review of Neuroimaging Techniques
Eight Popular Neuroimaging Techniques

- Structural magnetic resonance imaging (sMRI)
  - Diffusion weighted MRI (DWI)
  - Functional MRI (fMRI)
- Positron emission tomography (PET)
  - Computerized tomography (CT)
  - Electroencephalography (EEG)
  - Magnetoencephalography (MEG)
- Functional near-infrared spectroscopy (fNIRS)

Each image modality has its tracer, data dimension, extracted features, and main clinical and research applications.
CT, PET, MEG, EEG, and fNIRS

sMRI, fMRI, and DWI

- Gray matter (cortex + nuclei): cell bodies
- White matter: axons
- Myelin sheath speeds signal conduction
- Axon + sheath = nerve fibers
- Major white matter pathways aggregate many fibers into bundles

- Caudate
- Pallidum
- Putamen
- Amygdala
- Hippocampus
- Lateral Ventricle
- Thalamus
- Cortex
- White Matter

- Image showing brain structures and their corresponding labels.
A Multi-model Approach

- Different models at different scales.
- Ladder of overlapping models.
- Must be testable against multiple phenomena.
Part II

Imaging Processing Analysis Methods
Four Common Themes (CT1)-(CT4)

(CT1) Complex Brain Objects

- Complex Spatiotemporal Structures
  - spatial and temporal resolutions
  - spatio-temporal smoothness
  - spatiotemporal correlation

(CT2) Complex Spatiotemporal Structures

(CT3) Extremely High Dimensionality

(CT4) Heterogeneity within Individual Subjects and across Centers/Studies

https://dana.org/article/neuroanatomy-the-basics/
Image Models

Image = $f(B(age, gene, race, disease, others), device, acquisition, noises)$
Image Processing Analysis Methods

IPA: Deconvolution
- Image Reconstruction Process
- Image Enhancement Process

IPA: Structural Learning
- Image Segmentation Process
- Image Registration Process

Example: Airway Segmentation from CT
IPA: Deconvolution-IRP

Image Reconstruction Process (IRP)

\[
\frac{S(q)}{S_0} = \int P(r, \Delta) e^{i q \cdot r} dr; \quad q = \gamma \delta g u
\]
IPA: Deconvolution-IEP

Image Enhancement Process (IEP)
- Denoising
- Super-resolution
- Bias-field correction
- Harmonization
IPA: Structural Learning-ISP

Image Segmentation Process (ISP)
- Quantification of brain development
- Localization of pathology
- Surgical planning
- Image-guided interventions
- Computer-aided detection and diagnosis
- Brain parcellation

<table>
<thead>
<tr>
<th>Method</th>
<th>Dice_ET</th>
<th>Dice_WT</th>
<th>Dice_TC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase1</td>
<td>0.75245</td>
<td>0.89571</td>
<td>0.81561</td>
</tr>
<tr>
<td>Phase2 Model1</td>
<td>0.75983</td>
<td>0.90397</td>
<td>0.82489</td>
</tr>
<tr>
<td>Phase2 Model2</td>
<td>0.76091</td>
<td>0.90846</td>
<td>0.83622</td>
</tr>
<tr>
<td>Phase2 Model3</td>
<td>0.74669</td>
<td>0.89349</td>
<td>0.8278</td>
</tr>
<tr>
<td>Phase2 Model4</td>
<td>0.74187</td>
<td>0.90435</td>
<td>0.83211</td>
</tr>
<tr>
<td>Phase2 Model5</td>
<td>0.75779</td>
<td>0.90733</td>
<td>0.83824</td>
</tr>
<tr>
<td>Phase2 Model6</td>
<td>0.76091</td>
<td>0.90420</td>
<td>0.83713</td>
</tr>
<tr>
<td>Phase2 Model7</td>
<td>0.76814</td>
<td>0.90574</td>
<td>0.84704</td>
</tr>
<tr>
<td>Phase2 Model8</td>
<td>0.75440</td>
<td>0.90594</td>
<td>0.83826</td>
</tr>
<tr>
<td>Phase2 Model9</td>
<td>0.78582</td>
<td>0.90491</td>
<td>0.83689</td>
</tr>
<tr>
<td>XGBoost+</td>
<td>0.80506</td>
<td>0.91064</td>
<td>0.85037</td>
</tr>
</tbody>
</table>
IPA: Structural Learning-IRP

Image Registration Process (IRP)
- Automated image segmentation
- Construction of brain atlas
- Localization of pathology
- Multimodal fusion
- Population analysis
- Quantification of brain development
- Shape analysis
Brain Function-based Structural Connectome Atlas

Stage 1: Whole-brain Structural Connectome

Stage 2: Creation of Fiber Skeleton

Stage 3: Sparse Representation
Longitudinal Elastic Shape Data Analysis

Figure 2: A schematic
Challenges

There is no publicly available, high-quality neuroimaging datasets with detailed annotation information that cover a large spectrum of segmentation tasks in neuroimaging research.
Large-scale Neuroimaging related Studies
Large-scale Neuroimaging-related Studies

PING - 900 Pediatric Imaging, Neurocognition, and Genetics
BCP - 300 Baby Connectome Project
ADNI - 2000 Alzheimer’s Disease Neuroimaging Initiative
PNC - 1400 Philadelphia Neurodevelopmental Cohort
HCP - 1200 Human Connectome Project
ABCD - 10000 Adolescent Brain Cognitive Development
UKB - 500,000 UK Biobank Project
TCIA – 37,600 The Cancer Imaging Archive
NLST - 19,000 National Lung Screening Trial
OAI – 4800 Osteoarthritis Initiative
Alzheimer’s Disease Neuroimaging Initiative

The overall goal of ADNI is to validate potentially useful biomarkers for AD clinical treatment trials. ADNI is a multisite, prospective clinical study and actively supports the investigation and development of treatments that may slow or stop the progression of AD [https://adni.loni.usc.edu/study-design]. Researchers across 63 sites in the US and Canada have been tracking the progression of AD through clinical, imaging, genetic and biospecimen biomarkers, starting from normal aging, early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI) to dementia or AD.
The Human Connectome Project and Beyond

The primary goals of HCP include
• building a "network map" that will shed light on the anatomical and functional connectivity within the healthy human brain,
• promoting the understanding of inter-individual variability of brain circuits to behavior,
• facilitating research into brain disorders, such as autism, AD, and schizophrenia, and
• making all data freely available to the scientific community.

_The Heavily Connected Brain_
Peter Stern, “Connection, connection, connection…”, Science, Nov. 1 2013: Vol. 342 no. 6158 P.577
The ABCD study is the largest prospective longitudinal study of brain development and child health in the United States, which has recruited approximately 11,880 children aged 9-10 years old from 21 research sites and is following them for 10 years into early adulthood. Its initial goal was to examine risk and resiliency factors associated with the development of substance use, and then expanded far beyond, into identifying the underlying biospecimens, neural alterations, and environmental factors, and their contributions to the development of behavior, brain function, and other mental and physical outcomes throughout adolescence.

2015-now

https://abcdstudy.org/
The UK Biobank Study

UK Biobank has collected and continues to collect extensive environmental, lifestyle, and genetic data on half a million participants.

• **Imaging:** Brain, heart and full body MR imaging, plus full body DEXA scan of the bones and joints and an ultrasound of the carotid arteries. The goal is to image 100,000 participants, and to invite participants back for a repeat scan some years later.

• **Genetics:** Genotyping, whole exome sequencing & whole genome sequencing for all participants.

• **Health linkages:** Linkage to a wide range of electronic health-related records, including death, cancer, hospital admissions and primary care records.

• **Biomarkers:** Data on more than 30 key biochemistry markers from all participants, taken from samples collected at recruitment and the first repeat assessment.

• **Activity monitor:** Physical activity data over a 7-day period collected via a wrist-worn activity monitor for 100,000 participants plus a seasonal follow-up on a subset.

• **Online questionnaires:** Data on a range of exposures and health outcomes that are difficult to assess via routine health records, including diet, food preferences, work history, pain, cognitive function, digestive health and mental health.

• **Repeat baseline assessments:** A full baseline assessment is undertaken during the imaging assessment of 100,000 participants.

• **Samples:** Blood & urine was collected from all participants, and saliva for 100,000.
The major goals of ENIGMA include:

- Pushing forward the field of imaging genetics,
- Ensuring promising and reproducible findings,
- Sharing data, ideas, methods, algorithms and other information, and
- Training new investigators.

The Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) Consortium is a global alliance of over 1,400 scientists across 43 countries in the fields of imaging genomics, neurology, and psychiatry, studying a range of large-scale human brain studies that integrate data based on sMRI, DWI, fMRI, genetic data and many patient populations from over 70 institutions worldwide.

https://enigma.ini.usc.edu/
Part IV

Population-based Statistical Analysis Methods
Four Common Themes (CT5)-(CT8)

(CT5) Sampling Bias
- undercoverage
- observer bias,
- voluntary response bias
- survivorship bias
- recall bias
- exclusion bias

(CT6) Complex Missing Data Patterns
- missing by design
- faulty scanning
- attrition in longitudinal studies
- mis-entry
- non-responses in surveys

(CT7) Complex Data Objects

(CT8) Complicated Causal Pathways in Brain-related Disorders
Population-based Statistical Analysis (PSA)

❖ Study Design
❖ Statistical Parametric Mapping
❖ Object Oriented Data (OOD) Analysis
❖ Imputation Methods
❖ Data Integration Methods

➢ Dimension Reduction Methods
➢ Image Genetics
➢ Causality Research
➢ Predictive Analysis
➢ Knowledge-based Methods
Study Design

- Case control study
- Cross-sectional study
- Cohort study
- Experimental study
- Descriptive Study: case reports, case series, Descriptive surveys.

- The UKB is a large, population-based cohort study, and many cross-sectional analyses have been conducted based on baseline data from UKB.
- The UKB is well known for its "healthy volunteer" selection bias, and may not be a true representation of the general population.
- Neuroimaging biomarkers are usually secondary outcome.
**Statistical Parametric Mapping**

**Univariate Statistics**
- Preprocessed data: single voxel

**General linear model**
- Design matrix
- Parameter estimates

**SPMs**
- RFT/ permutation

**Multiple Comparisons**
- More complex models
- Multiple comparisons
Statistical Parametric Mapping

- From voxel-wise models to functional models
- Multiscale-adaptive estimation and inference procedures
- Wild-bootstrap methods to correct for multiple comparisons

\[ \text{Image} = f(B(\text{age, gene, race, disease, others}), \text{device, acquisition, noises}) \]
Parametric, Semiparametric and Nonparametric Models for OOD analyses

$$= g(x, \theta, f) \oplus \epsilon$$

$$x \in R^k, \theta \in \Theta \subset R^p, f \in F$$

$$g : R^k \times R^p \times F \rightarrow M$$
Intrinsic Regression Models

- **Feature Methods:** Use some feature extraction functions to project random objects to Euclidean-valued variables.

- **Extrinsic Methods:** Ignore the fact that manifold-valued data are in a nonlinear space and then directly apply classical multivariate regression.

- **Intrinsic Methods:** Few parametric models for manifold-valued data.

Intrinsic Regression Models

Geodesic Link Function

\[ g(x_i, \cdot) = q_0 + x_i q_1 = q_0 + x q_1 + (x_i - x) q_1 \]

Single-center

Fletcher (2013)

Maxwell et al. (2014)
Intrinsic Regression Models

How to define residual?

Inner product

Geodesic

Riemannian exponential maps

Riemannian logarithm maps

Residual

radius of injectivity
Intrinsic Regression Models

Conditional Mean

Riemannian logarithm maps

\[(x, q, b) = \text{Log}_{(x,q)}(Y) \quad T_{(x,q)}(Y)\]

Conditional Moment Model

\[E[(x,q)|x] = E[\text{Log}_{(x,q)}(Y)|x] = 0\]

Most major brain disorders (like AD) are heritable complex traits/diseases.

Together 50%-70% of AD risk
75%-90% of ADHD risk
60%-85% of Schizophrenia risk
~80% of Autism Spectrum Disorder (ASD) risk

Complex traits/diseases (many genes, environmental factors, complex functional mechanism)

Genetic signals are non-spare and weak:
Need large sample size to detect weak signals

Many genes contribute to the risk of AD (polygenic genetic architecture) (small but nonzero contribution)
Brain Imaging Genetics Knowledge Portal (BIG-KP)

Aim to build the best knowledge database of neuroimaging genetics

Heart Imaging Genetics Knowledge Portal (Heart-KP)
It's just a beginning

Hundreds of associated genetic variants for 2100+ neuroimaging traits across three modalities: (grey matter volume, white matter microstructure, resting-state functional connectivity+rfMRI, task fMRI, shape, heart)

1. Imaging Genetics Online Server

We build a GWAS browser using the Pheweb tool to explore GWAS results for massive functional, structural, and diffusion neuroimaging traits. Currently, we support GWAS results of 2104 traits trained in the UK British cohort (n=34,000), including

- 635 ENIGMA-DTI parameters of brain white matter (diffusion MRI)
- 376 ANTS regional brain volumes (structural MRI)
- 191 ICA-based functional MRI traits (rs-fMRI,ICA))
- 760 parcellation-based functional MRI traits (task+fMRI,classifier260)
For a heterogeneous, clinically defined disorder, the endophenotype is ‘closer to the underlying biology,’

- Increasing the power of genetic search
- Being informative about disorder risk.
- Providing mechanistic connections linking genetic variation to behavioral measures.
Model Setup

Outcome generating model

\[ Y_i = \sum_{l=1}^{s} x_{il} \beta_l + < Z_i, B > + \epsilon_i \]

Exposure generating model

\[ Z_i = \sum_{l=1}^{s} x_{il} \ast C_l + E_i \]

\( B \) is the main parameter of interest, representing the association between the 2D imaging exposure \( Z_i \) and the behavioral outcome \( Y_i \), \( \beta_l \) represents the association between the \( l \)-th observed covariate \( x_{il} \) and the behavioral outcome \( Y_i \), and \( \epsilon_i \) and \( E_i \) are random errors that may be correlated. The symbol “\( \ast \)” denotes element-wise multiplication.

True Confounders, Precision, Instrumental and Irrelevant Variables

Outcome generating model
\[ Y_i = \sum_{l=1}^{s} x_{il} \beta_l + < Z_i, B > + \epsilon_i \]

Exposure generating model
\[ Z_i = \sum_{l=1}^{s} x_{il} \ast C_l + E_i \]

**True Confounders**
\[ C = \{ l \in A | \beta_l \neq 0 \text{ and } C_l \neq 0 \} \]

**Precision Variables**
\[ P = \{ l \in A | \beta_l \neq 0 \text{ and } C_l = 0 \} \]

**Instrumental Variables**
\[ I = \{ l \in A | \beta_l = 0 \text{ and } C_l \neq 0 \} \]

**Irrelevant Variables**
\[ S = \{ l \in A | \beta_l = 0 \text{ and } C_l = 0 \} \]

Aim (to correctly estimate \( B \)): retain all covariates from \( M_1 = C \cup P = \{ l \in A | \beta_l \neq 0 \} \), while excluding covariates from \( I \cup S = \{ l \in A | \beta_l = 0 \} \).
Marginal Screening

Fit:
\[ Y_i = x_{il} \beta_l + \epsilon_i \]

Obtain:
\[ \hat{\beta}_l^M = n^{-1} \sum_{i=1}^n x_{il} Y_i \]

Problem!!! (plugging exposure model into outcome model)

Outcome generating model
\[ Y_i = \sum_{l=1}^s x_{il} \beta_l + < Z_i, B > + \epsilon_i \]

Exposure generating model
\[ Z_i = \sum_{l=1}^s x_{il} * C_l + E_i \]

Obtain:
\[ Y_i = \sum_{l=1}^s x_{il} (\beta_l + < C_l, B >) + < E_i, B > + \epsilon_i \]

Miss a portion of confounders when \( \beta_l \) and \( < C_l, B > \) are of similar magnitude but opposite sign.
Joint Screening (proposed)

Marginal screening:

$$Z_i = \sum_{l=1}^{s} x_{il} \ast C_l + E_i$$

Obtain (Kong, An, Zhang and Zhu, 2020):

$$\hat{C}_l^M = n^{-1} \sum_{i=1}^{n} x_{il} \ast Z_i \in \mathbb{R}^{p \times q}$$

$$\hat{M}_1^* = \left\{ 1 \leq l \leq s: \left| \beta_l^M \right| \geq \gamma_{1,n} \right\}$$

$$\hat{M}_2 = \left\{ 1 \leq l \leq s: \| \hat{C}_l^M \|_{op} \geq \gamma_{2,n} \right\}$$

Select submodel: $\hat{M} = \hat{M}_1^* \cup \hat{M}_2.$ (Union)

Alternative choices (both worse): $\hat{M}_1^*$ (outcome) or $\hat{M}_1^* \cap \hat{M}_2$ (Outcome).
Estimation (proposed)

Minimize:

\[
\frac{1}{2} \sum_{i=1}^{n} \left( Y_i - \langle Z_i, B \rangle - \sum_{l \in \hat{M}} X_{il} \beta_l \right)^2 + \lambda_1, n \sum_{l \in \hat{M}} |\beta_l| + \lambda_2, n \| B \|_*
\]

where \( \| B \|_* = \sum_k \sigma_k(B) \).

L1 penalty, exclude instrumental and irrelevant variables.

Nuclear penalty, low-rank estimation of \( B \).

Estimated effect size of imaging exposure \( z \),

\[
\hat{\mu}(z) = \langle z, \hat{B} \rangle
\]
Gene expression-informed gene-level PRS + GWAS PRS has higher prediction accuracy

Construct gene-level PRS (polygenic risk scores) by leveraging gene expression reference panels (e.g., GTEx) in TWAS
Causal Genetics Imaging Clinical Pathway

Environment

Genomics

Imaging

Disease

Cognition

Behavior

Unseen Confounder

Confounder

UNC

Biostatistics
Brain Imaging Genetics Paradigm

Neuroimaging: an important component to help understand the complex biological pathways of brain disorders.

Genes
- RNA genes, protein-coding genes
- Genomics
- Epigenomics
- Gene expression at RNA and protein levels
- Epigenomic modifications
c
- De novo mutations

Molecules
- RNA, proteins, metabolites
- Transcriptomics
- Proteomics
- Metabolomics
- Interactomics
- Molecular function and cell metabolism

Cells
- Neuron development, organelle
- Cell biology
- Neuroscience

Brain
- Structure, circuits, physiology
- Changes in neural interactions, altered brain structure/function
- Uncover the profile of brain abnormalities in each clinical outcome to study how disorders develop

Symptoms
- Behavioral tests

Biological causes
- Molecular function and cell metabolism

Environmental, social and psychological factors
- Social and psychological influences

BIG-KP | https://bigkp.org
Challenges

❖ The complexity of those large-scale neuroimaging-related data sets is too high for most research teams in both academia and industry.

❖ It is very difficult to appropriately process data across different domains with high quality, while controlling for potential bias introduced during the preprocessing stage.

❖ It remains uncertain as to how to appropriately integrate data across different domains obtained from different studies and cohorts with possible different study designs for unbiased data integration.

❖ It remains unclear how to appropriately and efficiently analyze neuroimaging related data sets with multiple Vs (e.g., Volume, Velocity, Variety and Veracity), while ensuring algorithmic fairness.
Statistical Learning Methods for Neuroimaging Data Analysis with Applications

Hongni Zhu1, Xingli Li2, and Bingxin Zhao3

1Department of Biostatistics, Statistics, and Computer Science and Biomedical Research Imaging Center, University of North Carolina, Chapel Hill, NC; 27514 (email: jzhou@bios.unc.edu)
2Department of Radiology and Biomedical Research Imaging Center, University of North Carolina, Chapel Hill, NC; 27514
3Department of Radiology and Biostatistics, University of Pennsylvania, Philadelphia, PA; 19104

Keywords: causal pathway, image processing, neuroimaging, analysis, statistical techniques, neuroimaging analysis, study design

Abstract

The aim of this paper is to provide a comprehensive review of statistical challenges in neuroimaging data analysis, along with recent techniques and their potential applications in neuroimaging research and clinical translation. We review the four core themes of neuroimaging data analysis and present major image processing analysis methods for processing neuroimaging data at the individual level. We briefly review the large-scale neuroimaging natural studies and a collection of imaging comments and discuss several common themes of neuroimaging data analyses in the population level. We review major population-based statistical analysis methods and their associated statistical challenges and present recent progress in statistical methodology to address these challenges.

(Zhu, Li & Zhao, 2023)
Acknowledgement

**Funding:** U.S. NIH Grants MH086633 and MH116527

**Pictures:** Copyrights belong to their own authors and/or holders.

**Data:** We thank Bingxin Zhao, Tengfei Li and other members of the UNC BIG-S2 lab (https://med.unc.edu/bigs2/) for processing the neuroimaging data.

UK Biobank resource application number: 22783.