Statistical Methods for Neuroimaging Data Analysis
ICSA Canada Chapter 2015 Lecture 4
Imaging Genetics

Linglong Kong
University of Alberta

ICSA Canada Chapter August 4, 2015 @ Calgary, AB

Special Thanks to Professor Hongtu Zhu at UNC
References


Motivation
Imaging genetics allows for the identification of how common/rare genetic polymorphisms influencing molecular processes (e.g., serotonin signaling), bias neural pathways (e.g., amygdala reactivity), mediating individual differences in complex behavioral processes (e.g., trait anxiety) related to disease risk in response to environmental adversity.

Multivariate, smoothed functions, and piecewisely smoothed functions
Dimension varies from 100~500,000.
Brain Structure is highly heritable

Must be specific genetic variants explaining the high heritability (Kremen et al.2010).
Genetic Data
Features:
-- Spatially correlated functional data
-- High dimension (~ $2^{21} = 2,097,152$ voxels)
-- Voxels with similar color form regions
Genetic Variation Data

- S NP (Single Nucleotide Polymorphism)
- Correlated
- High Dimension
Imaging Genetic Studies -- Brain Development

http://www.brain-map.org/

UNIVERSITY of ALBERTA
Imaging Genetic Studies -- Brain Tumor

A

Microarray Samples
- GBM
- Normal Brain

MRI Phenotypes
- Contrast / Necrosis
- Contrast Enhancement
- Infiltrative
- Mass Effect
- SVZ involvement

B

Contrast enhancement
- High
- Low

C

Mass effect
- High
- Low
Imaging Genetic Studies
-- More Studies
Genetic variation through collaboration

- Common genetic variation has only a **small effect** on brain structure.

- In a GWAS, many statistical tests are conducted.

- **Huge sample sizes** (>10,000 individuals) are needed in order to find and replicate association of individual genetic variants.

- As both imaging and genotyping are expensive, a **consortium is needed**.

  Dr. Hibar

http://enigma.ini.usc.edu

> 185 institutions, 300+ co-authors, world-wide consortium countries resulting in ~30,000+ subjects

*Now expanded to multiple multi-national meta-analytic projects:*

- Genetics of brain structure
- Genetics of brain function
- Meta-analytic effects on disease
Procedure for meta-analysis (Part 1)

All protocols available at http://enigma.ini.usc.edu
Procedure for meta-analysis (Part 2)

Genome-wide association to imaging phenotypes (accounting for kinship in related samples)
(50 contributing sites with maximum N = 29,556)

Phenotypes (volume):
- ICV
- Nucleus Accumbens
- Amygdala
- Caudate
- Hippocampus
- Pallidum
- Putamen
- Thalamus

Covariates:
- ICV (for non-ICV phenotypes)
- Age
- Sex
- Age-squared
- 4 MDS
- Dummy covariate for scanner
- Patient status

Quality Checking and Filtering (MAC < 10, R² < 0.5)

Meta-analyses
Fixed-effect, inverse variance-weighted model

Manhattan Plot (Putamen Phenotype)

QQ Plot (Putamen Phenotype)

Lambda = 1.006
ENIGMA1: Pilot Project Hippocampal and Intracranial Volume GWAS Meta-analysis

21,151 individuals in discovery + replication

(Stein et al., Nature Genetics, 2012)
ENIGMA2: Genetics of subcortical structures

Discovery + Replication = 30,717 individuals
50 contributing cohorts around the world

(Hibar et al., 2015. Nature)
What’s next for ENIGMA?

ENIGMA3: GWAS of cortical area and thickness
Projected N=30,000
Interested in joining? Sign up at http://enigma.ini.usc.edu/
ENIGMA-DTI

Meta-analysis (N)

Meta-analysis (SE)

Mega-analysis

Meta-analysis (N)

Meta-analysis (SE)

Mega-analysis

Individual Site Heritability

http://enigma.ini.usc.edu/ongoing/dti-working-group/
Mapping Connectivity

- Structural magnetic resonance imaging (MRI)
  - Anatomical analysis
  - Mapping Tissue Loss

- Diffusion weighted
  - Microstructure
  - Connectivity
Connectome-Wide Genome-Wide Screen Alzheimer risk gene

Discovery sample – Young Adults
Effect in ADNI
Within 2 weeks Sherva et al. published *SPON1*
Found in a cognitive GWAS in AD

Jahanshad et al., PNAS 2013
Overview
Imaging Genetic Studies

Common Goal: Detect potential genes for inherited phenotypes
<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="image" alt="Candidate SNP" /></td>
<td><img src="image" alt="Imager" /></td>
<td><img src="image" alt="Imager" /></td>
</tr>
<tr>
<td>Candidate SNP</td>
<td><img src="image" alt="Geneticist" /></td>
<td><img src="image" alt="Geneticist" /></td>
<td><img src="image" alt="Geneticist" /></td>
</tr>
<tr>
<td>Candidate Gene</td>
<td><img src="image" alt="Geneticist" /></td>
<td><img src="image" alt="Geneticist" /></td>
<td><img src="image" alt="Geneticist" /></td>
</tr>
<tr>
<td>Genome-wide SNP</td>
<td><img src="image" alt="Geneticist" /></td>
<td><img src="image" alt="Geneticist" /></td>
<td><img src="image" alt="Geneticist" /></td>
</tr>
<tr>
<td>Genome-wide Gene</td>
<td><img src="image" alt="Geneticist" /></td>
<td><img src="image" alt="Geneticist" /></td>
<td><img src="image" alt="Geneticist" /></td>
</tr>
</tbody>
</table>

Hibar, et al. HBM 2012
High Dimensional Regression Model

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

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

Phenotype \( Y \)

Genotype \( X \)

Error \( E \)

\( Y \times p_y = X + B + E \)

\( n \times p_y \)

\( n \times p_x \)

\( p_x \times p_y \)

\( n \times p_y \)

\( (p_x, n, p_y) \)

UNIVERSITY of ALBERTA
Current Approaches and Limitations

- Candidate Gene And / Or Candidate Imaging Region
- Genomic-wide Search + Voxel-based Approach
- Everything in a big model + Penalized Method

Only part of the picture

- correlation (LD) structure
- spatial correlation in voxels
- Low statistical power
- Unstable variable selection + Statistical inference

http://www.tnooz.com/
Models for Candidate ROI Candidate Gene-set Analysis
High Dimensional Regression Model

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

\[ Y_i = \{y_i(v) : v \in V_0 \} \quad X_i = \{X_i(g) : g \in G_0\} \]

Key Conditions:
- Sparsity of \( B \)
- Restricted null-space property for design matrix \( X \)

\[ \text{max}(p_x, p_y) \sim n \]
Sparse and Low-rank Representation

Sparsity on B.

\[ B \times p_x \times p_y \]

Low Rank

\[ b_X \]

Sparsity

\[ b_Y \]

Regularization Methods

- Lasso 1, 2, 3, ....
- SCAD, MCP, ....

\[ \hat{\theta} \in \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda_n \sum_{j=1}^{p} |\theta_j| \]
Vounou et al. (2010, 2012) NeuroImage
Group all available SNPs into $L$ pathways $\mathcal{G}_1, \ldots, \mathcal{G}_L$

Adopt a group lasso penalty to force group selection,

$$P_b(b) = \frac{1}{2}||y - Xb||^2_2 + \lambda \sum_{l=1}^{L} w_l ||b_l||^2_2$$

to select only the most predictive pathways,

$$b = \{(0, \ldots, 0), \ldots, (0, \ldots, b_{i_b}, 0 \ldots, b_{i_b}, 0, \ldots, 0), \ldots, (0, \ldots, 0)\}$$
Overlapping Pathways

SNPs Duplication in Overlapping Pathways

Three pathways $g_1, g_2, g_3$ and grouped regression coefficients $\beta_1, \beta_2, \beta_3$
Serial brain MRI scans were analyzed from 200 probable AD patients and 232 healthy elderly controls (CN).

Longitudinal scans available at three time points

<table>
<thead>
<tr>
<th></th>
<th>Screening</th>
<th>6Mo</th>
<th>12Mo</th>
<th>24Mo</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>200</td>
<td>165</td>
<td>144</td>
<td>111</td>
</tr>
<tr>
<td>CN</td>
<td>232</td>
<td>214</td>
<td>202</td>
<td>178</td>
</tr>
<tr>
<td>Total</td>
<td>432</td>
<td>379</td>
<td>346</td>
<td>289</td>
</tr>
</tbody>
</table>

At screening:

<table>
<thead>
<tr>
<th>Group</th>
<th>age (years)</th>
<th>N male</th>
<th>N female</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>75.7±7.7</td>
<td>103</td>
<td>97</td>
</tr>
<tr>
<td>CN</td>
<td>76.0±5.0</td>
<td>120</td>
<td>112</td>
</tr>
</tbody>
</table>
AD Imaging Signature

Mean and STD of slopes

-log10(p) value map
Mapping SNPs to Pathway

ADNI QC’d SNPs
434,271 SNPs

Genes: GRCH36/hg18
21,004 genes

SNP to gene mapping
211,106 SNPs mapped to
18,405 genes within 10kbp

SNP to pathway mapping

Pathways: KEGG
186 Pathways containing
5,267 distinct genes

Exclude largest, highly redundant, pathway

66,162 SNPs mapped to 4,632 genes and 185 pathways

Overlap expansion

$P^* = 175,544$ SNPs mapped to 185 pathways
## Top 15 SNPs and Genes Ranked by SRRR

<table>
<thead>
<tr>
<th>Rank</th>
<th>SNP</th>
<th>$\pi^{SNP}$</th>
<th>Mapped gene(s)</th>
<th>Gene</th>
<th>$\pi^{gene}$</th>
<th># SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs4788426</td>
<td>0.451</td>
<td>PRKCB</td>
<td>PRKCB</td>
<td>0.451</td>
<td>73</td>
</tr>
<tr>
<td>2</td>
<td>rs11074601</td>
<td>0.429</td>
<td>PRKCB</td>
<td>ADCY8</td>
<td>0.411</td>
<td>69</td>
</tr>
<tr>
<td>3</td>
<td>rs263264</td>
<td>0.411</td>
<td>ADCY8</td>
<td>ADCY2</td>
<td>0.392</td>
<td>106</td>
</tr>
<tr>
<td>4</td>
<td>rs13189711</td>
<td>0.392</td>
<td>ADCY2</td>
<td>HK2</td>
<td>0.302</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>rs680545</td>
<td>0.302</td>
<td>HK2</td>
<td>PRKCA</td>
<td>0.290</td>
<td>99</td>
</tr>
<tr>
<td>6</td>
<td>rs4622543</td>
<td>0.290</td>
<td>PRKCA</td>
<td>PIK3R3</td>
<td>0.267</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>rs9806483</td>
<td>0.274</td>
<td>PRKCA</td>
<td>MYLK</td>
<td>0.234</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>rs1052610</td>
<td>0.267</td>
<td>PIK3R3</td>
<td>PIK3CG</td>
<td>0.207</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>APOe4</td>
<td>0.251</td>
<td>TOMM40 APOE</td>
<td>COL5A3</td>
<td>0.174</td>
<td>14</td>
</tr>
<tr>
<td>10</td>
<td>rs1254403</td>
<td>0.234</td>
<td>MYLK</td>
<td>GNAI1</td>
<td>0.167</td>
<td>22</td>
</tr>
<tr>
<td>11</td>
<td>rs4730205</td>
<td>0.207</td>
<td>PIK3CG</td>
<td>ACACA</td>
<td>0.164</td>
<td>23</td>
</tr>
<tr>
<td>12</td>
<td>rs889130</td>
<td>0.174</td>
<td>COL5A3</td>
<td>G6PC</td>
<td>0.163</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>rs6973616</td>
<td>0.167</td>
<td>GNAI1</td>
<td>DGKA</td>
<td>0.160</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>rs9906543</td>
<td>0.164</td>
<td>ACACA</td>
<td>CR1</td>
<td>0.154</td>
<td>21</td>
</tr>
<tr>
<td>15</td>
<td>rs2229611</td>
<td>0.163</td>
<td>G6PC</td>
<td>TOMM40</td>
<td>0.152</td>
<td>6</td>
</tr>
</tbody>
</table>
Factor Model

\[ E \]
\[ n \times p_y \]

Long-range Correlation

\[ E_i \]
\[ p_y \times 1 \]

\[ \Lambda \]
\[ p_y \times q \]
\[ q \times 1 \]

\[ \xi_i \]
\[ \eta_i \]

Short-range Correlation

\[ p_y \times 1 \]

\[ \Sigma_E \]

\[ = \]

\[ \Lambda^{T} \]

\[ \Sigma \eta \]
Simulation

<table>
<thead>
<tr>
<th>Patterns</th>
<th>Plus</th>
<th>SVD</th>
<th>SVD</th>
<th>UN</th>
<th>UN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True B</strong></td>
<td><img src="image1" alt="Pattern" /></td>
<td><img src="image2" alt="Pattern" /></td>
<td><img src="image3" alt="Pattern" /></td>
<td><img src="image4" alt="Pattern" /></td>
<td><img src="image5" alt="Pattern" /></td>
</tr>
<tr>
<td><strong>LASSO</strong></td>
<td><img src="image6" alt="Pattern" /></td>
<td><img src="image7" alt="Pattern" /></td>
<td><img src="image8" alt="Pattern" /></td>
<td><img src="image9" alt="Pattern" /></td>
<td><img src="image10" alt="Pattern" /></td>
</tr>
<tr>
<td><strong>BLASSO</strong></td>
<td><img src="image11" alt="Pattern" /></td>
<td><img src="image12" alt="Pattern" /></td>
<td><img src="image13" alt="Pattern" /></td>
<td><img src="image14" alt="Pattern" /></td>
<td><img src="image15" alt="Pattern" /></td>
</tr>
<tr>
<td><strong>G-SMuRFS</strong></td>
<td><img src="image16" alt="Pattern" /></td>
<td><img src="image17" alt="Pattern" /></td>
<td><img src="image18" alt="Pattern" /></td>
<td><img src="image19" alt="Pattern" /></td>
<td><img src="image20" alt="Pattern" /></td>
</tr>
<tr>
<td><strong>GLRR3</strong></td>
<td><img src="image21" alt="Pattern" /></td>
<td><img src="image22" alt="Pattern" /></td>
<td><img src="image23" alt="Pattern" /></td>
<td><img src="image24" alt="Pattern" /></td>
<td><img src="image25" alt="Pattern" /></td>
</tr>
<tr>
<td><strong>GLRR5</strong></td>
<td><img src="image26" alt="Pattern" /></td>
<td><img src="image27" alt="Pattern" /></td>
<td><img src="image28" alt="Pattern" /></td>
<td><img src="image29" alt="Pattern" /></td>
<td><img src="image30" alt="Pattern" /></td>
</tr>
</tbody>
</table>

UNIVERSITY of ALBERTA
749 AD/MCI/NC subjects, 93 ROIs
40 AD candidate genes on the AlzGene web
Figure 4: Results of ADNI data: the posterior estimate of $\hat{B}$ matrix after thresholding out elements whose $p$-values are greater than 0.001 (left panel), $B^T_{\text{bin}} B_{\text{bin}}$ (middle panel) and $B^T_{\text{bin}} B_{\text{bin}}$ (right panel) in the first row; and the log$_{10}$ $p$-value matrices corresponding to $B$ (left panel), $U$ (middle panel), and $V$ (right panel) in the second row.

Figure 5: Results of ADNI data: the top 20 ROIs based on $B^T_{\text{bin}} B_{\text{bin}}$ and the first 3 columns of $V$. The sizes of the dots represent the rank of the ROIs.

$-\log_{10}(p)$ for $\hat{B}$
Models for Candidate SNP/Gene Voxel-wise Analysis
Candidate SNP

rs11136000 (CLU)

Genome-wide association identifies variant within the CLU gene in ~4000 Alzheimer’s patients and ~8000 controls – but what does it do?

(Harold et al., 2009)

The Alzheimer’s associated variant broadly affects white matter integrity in a young cohort – may create early predisposition for disease

(Braskie et al., 2009)
High Dimensional Regression Model

Data \[ \{(Y_i, X_i) : i = 1, \ldots, n\} \]

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

Phenotype \( Y \)

Genotype \( X \)

Error \( E \)

\[ n \times p_y = n \times p_x + p_x \times p_y + n \times p_y \]

\( p_x \ll n \ll p_y \)
Functional Mixed Effects Model

\[ y_i(v) = x_i' \beta(v) + z_i' \gamma(v) + e_i(v) \]

\[ \gamma(v) \sim N(0, \sigma^2_{\gamma}(v) I_L) \]
\[ e_i(v) \sim N(0, \sigma^2_e(v)) \]
\[ \dim(y_i(v)) = 1 \]

Imaging Responses

Genetic Variation

\[ \sigma^2_{\gamma}(v) = 0 \]

UNIVERSITY of ALBERTA
Estimate $\sigma^2(\nu)$, $\sigma^2(\nu)$ marginally, using regular REML

Estimate $\sigma^2(\nu)$ adaptively, using weighted REML

Hypothesis Testing for $H_0(\nu)$: $\sigma^2(\nu) = 0$
Weighted REML

\[ Y_{nx1}(v) = X_{nxp}\beta(v)_{px1} + Z_{nxL}\gamma(v)_{Lx1} + E(v)_{nx1} \]

\[ K_{(n-p)xn} X_{nxp} = 0 \]

\[ Y^*(v) = KY(v) = K Z\gamma(v) + K E(v) = Z^* \gamma(v) + E^*(v) \]

Regular REML

Weighted REML

\[ B(v,h) : \]
Hypothesis Testing

H₀ : \( \sigma^2_\gamma(v) = 0 \) vs H₁ : \( \sigma^2_\gamma(v) \neq 0 \)

Test statistic \( T(v) : \)

\[
2\{L_{REML}(Y^*(v) | Z, \hat{\sigma}^2_\gamma(v), \hat{\sigma}^2_e(v)) - L_{REML}(Y^*(v) | Z, 0, \hat{\sigma}^2_e(v))\}
\]

Traditional approach: \( T \sim \frac{1}{2} \chi^2_0 + \frac{1}{2} \chi^2_1 \)

Our approach : Exact Null Distribution

\[
T_n(v) \sim \sup_{\sigma^2(v) \geq 0} \sum_{B(v,h)} \omega(v,v',h) \sum_{i=1}^{L} \frac{\eta_i^2(v) \sigma^2_\gamma(v)d_i}{\hat{\sigma}^2_e(v') + \sigma^2_\gamma(v)d_i} - \log(\hat{\sigma}^2_e(v')) + \sigma^2_\gamma(v)d_i
\]

\( d_i : i^{th} \) eigenvalue of \( KZ^TZK^T \)

\( \eta_i(v) \sim N(0,1) \)
Score Test

Likelihood Ratio Test:
-- Smoothing estimator then computing LRT
-- Solve it for EVERY random sample and EVERY voxels

Regular Score Test: \( SC(v) = \frac{1}{2} \sigma_{ev}^{-4} Y_v^T \Omega Y_v \) (Tzeng2007)

-- No need to solve maximizer every time

-- \( E(SC) \) and \( E(SC^2) \) feasible to calculate

But statistical power is not good enough…
Weighted Score Test

Regular Score Test:

\[ SC_{\sigma^2(v)} = \frac{1}{2} \sigma_{\text{ev}}^{-4} Y_v^T \Omega Y_v \sim \text{Gamma}(\alpha, \beta) \]

(Tzeng 2007)

Weighted Score Test:

\[ WSC_{\sigma^2(v)} = \frac{1}{2} \sum_{v' \in B(v, h)} \omega_{v'} \sigma_{\text{ev'}}^{-4} Y_{v'}^T \Omega Y_{v'} \sim \text{Gamma}(\alpha', \beta') \]

No tricks: 20 Mins!
Results: Hypothesis Testing – Local view

\[ H_0 : \sigma^2_{\gamma(v)} = 0 \ vs \ H_1 : \sigma^2_{\gamma(v)} \neq 0 \]

Voxel Based

Score T

LRT

True Value

FMEM
Objective: Identify brain regions affected by CR1

Subject: 335 elders (174 Controls, 161 ADs)

Image: 128 * 128 * 128 RAVEN maps from baseline T1-weighted images

Demographic Info: Gender, Baseline age, Baseline age square, APOE Risk, Handedness, Education Level, Baseline intracranial volume

Genetic Variation: CR1 (16 SNPs, MAFs from 13% ~ 41%)

Method: Functional Mixed Effects Model (FEME)
45 ROIs by FMEM, 5 ROIs by Voxel-based

Superior temporal gyrus (R167)
Inferior temporal gyrus (R150, L378)
Precentral gyrus (R170, L*439)
Middle frontal gyrus (R127, L295)
Postcentral gyrus (R60, **L65)
Insula (R53, L74)
Putamen (R200, L*123)
Pallidum (R58)
Fusiform (R117, L306)
Inferior temporal gyrus (R219)
Inferior parietal but supramarginal and angular gyri (R87)
Angular (R433)
Inferior frontal gyrus, triangular part (R180, L*79)
Inferior occipital gyrus (R269, L133)
Superior frontal gyrus (R71)
Supplementary Motor Area (R117)
Postcentral gyrus (R104, L183)
Superior frontal gyrus, medial (R188, L159)
Anterior cingulate and paracingulate gyri (R428, L88)

Median cingulate and paracingulate gyri (R209, L116)
Calcarine fissure and surrounding cortex (R53, L266)
Cuneus (R151, L345)
Superior occipital gyrus (L317)
Middle occipital gyrus (R144)
Precuneus (R61)
Paracentral Lobule (R86)
Caudate (R197)
Lingual (L114)
Inferior frontal gyrus, opercular part (L78)
Middle temporal gyrus (L*394)
Inferior frontal gyrus, orbital part (L61)
Models for Candidate ROIs Whole Genome-wise Analysis
High Dimensional Regression Model

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

\[ Y_i = \{y_i(v) : v \in V\} \]

\[ \{X_i(g) : g \in G_0\} \]

Phenotype \( Y \)

Genotype \( X \)

Error \( E \)

\( n \times p_y \)

\( n \times p_x \)

\( p_x \times p_y \)

\( n \times p_y \)

\( p_x \ll n < p_y \)
Current Approaches

Multivariate Responses

- Multivariate Linear Model + Hotelling’s $T^2$ Test
- Principle Component Regression
- Component Wise Method (False Discovery Rate)
- Partial Least Square Regression
Limitations

- **Multivariate Linear Model + Hotelling’s $T^2$ Test**
- **Principle Component Regression**
- **Component Wise Method**
- **Partial Least Square Regression**

- **N > p Hotelling $T^2$ Test**
- **Low Statistical Power**
- **Inflated Type I Error Rate**
- **No p-values!!**

- **Larger N**
- **Wrong Conclusions**
Sparse Projection Regression Model

- Multivariate regression with a high-dimensional responses and a multivariate covariate of interest
- Consider a Multivariate Linear Model (MLM):

  \[ Y = XB + E, \quad \text{or} \quad y_i = B^T x_i + e_i \]

- We are interested in the hypothesis testing problem:

  \[ H_0 : CB = B_0 \quad \text{v.s.} \quad H_1 : CB \neq B_0 \]

- Diverging \( q \), fixed \( p \) case
  - High-dimension two sample test
  - Imaging genetics association study
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 $b_0 = B_0 W$

How to determine an 'optimal' $W$?
Sparse Projection Regression Model

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

$$GHR(w; C) = \frac{w^T(\tilde{B}_1 - B_0)^T S_{\tilde{X}_1}(\tilde{B}_1 - B_0)w}{w^T \Sigma R w} = \frac{w^T \Sigma C w}{w^T \Sigma R w}$$

- High Dimensional Setting
- noise accumulation
  - ill-conditioned sample covariance estimator: \( \hat{\Sigma}_R \)
- Sparse Projection Regression Model is proposed as following:

$$\arg\max\left\{ \frac{w^T \hat{\Sigma}_C w}{w^T \hat{\Sigma}_R w} \right\} \quad \text{s.t.} \quad \|w\|_1 \leq t$$
After estimating $W$, we can calculate a $k \times k$ matrix as:

$$T_n = (C\hat{\beta}_w - b_0)^T \Sigma^{-1}_\Omega (C\hat{\beta}_w - b_0)$$

- Test statistic: $Tr_n = \text{trace}(T_n)$
- Wild bootstrap
  - Fit MLM under the null hypothesis to calculate the estimated multivariate regression coefficient, denoted by $\hat{B}_0$, residuals $\hat{e}_i = y_i - \hat{B}_0^T x_i$.
  - Generate $G$ bootstrap samples $z_i^{(g)} = (\hat{B}_0)^T x_i + \eta_i^{(g)} \hat{e}_i$.
  - Repeat the estimation procedure for estimating the optimal weights and the calculation of the test statistic $Tr_n^{(g)}$.
  - $p$-value of $Tr_n$ is computed as $\frac{1}{G} \sum_{g=1}^{G} 1(Tr_n^{(g)} \geq Tr_n)$.
Numerical Example: High Dimensional Two Sample Test

- \{y_1, \ldots, y_{n_1}\} and \{y_{n_1+1}, \ldots, y_n\} \subset \mathbb{R}^q from N(\beta_1, \Sigma_R) and N(\beta_2, \Sigma_R), respectively.
- We set: \(n = 2n_1 = 100\) and \(q\) is 50, 100, 200, 400, 800, 1000, 1500, and 2000, respectively.
- \(H_0 : \beta_1 = \beta_2\) against \(H_1 : \beta_1 \neq \beta_2\)
- Can be formulated by a regression model with \(B^T = [\beta_1, \beta_2]\) and \(C = (1, -1)\).
- Error covariance matrix \(\Sigma_R = \sigma^2(\rho_{j,j'})\):
  - Model 1: is an independent covariance matrix with \((\rho_{jj'}) = \text{diag}(1, \cdots, 1)\).
  - Model 2: is a weak correlation matrix with 
    \[\rho_{jj'} = 1(j' = j) + 0.3 \times 1(j' \neq j).\]
  - Model 3: is a strong correlation covariance matrix with \(\rho_{jj'} = 0.8^{|j' - j|}\).
Simulation

---

Power

![Graph 1](image1.png)

![Graph 2](image2.png)

Type I error

![Graph 3](image3.png)

![Graph 4](image4.png)
Alzheimer’s Disease Neuroimaging Initiative (ADNI) Data Analysis

- Problem of interest: to perform a genome-wide search for establishing the association between the 10,479 SNPs collected on chromosome 19 and the brain volume of 93 regions of interest (ROIs).
- 93 ROIs—intercept, a specific SNP, age, gender, whole brain volume, the top 5 principal components
- Several important findings:
  - The ApoE allele was identified as the top one significant covariate with $-\log_{10}(p) \sim 15$.
  - The rs207650 on the TOMM40 gene is found to be one of the top 3 significant SNPs with $-\log_{10}(p) \sim 5$, Vounou et al. (2012).
  - Able to detect some additional SNPs, such as rs11667587 on the NOVA2 gene.
Fast Voxel-wise Genome-wise Analysis
Data Structure

Imaging:

Genetic Variation:

Person No. 1

\[
\begin{bmatrix}
1 & 2 & \cdots & 0 \\
0 & \ddots & & 1 \\
\vdots & & \ddots & \vdots \\
1 & 0 & \cdots & 2
\end{bmatrix}
\]
<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate SNPs allow you to test a specific biological hypothesis</td>
<td>It is highly likely that we don’t know the genetic underpinnings of a trait like brain structure so in general don’t know the right SNP to pick</td>
</tr>
<tr>
<td>Strong hypothesis drives clearly interpretable results</td>
<td>In order to be widely accepted, the variant needs to have strong prior evidence (genome-wide significant in a meta-analysis or have clear function)</td>
</tr>
<tr>
<td>Multiple comparisons burden is reduced (one SNP – many voxels)</td>
<td>Unable to search the genome, only characterize the effect of a known variant</td>
</tr>
<tr>
<td>Quick way to provide functional relevance to unbiased genome-wide search results</td>
<td></td>
</tr>
</tbody>
</table>
vGWAS

~1000 subjects

~30,000 voxels in the brain

~600,000 genetic markers (SNPs)

1.8 x 10^{10} tests!

Hibar, et al. HBM 2012
Figure 1: Summary of vGWAS method for conducting association and assessing statistical significance.

Stein et al. 2010
Raw Minimum p-value at each voxel

Hibar et al. 2011
## Most associated genes

<table>
<thead>
<tr>
<th>Chr</th>
<th>Gene</th>
<th># of SNPs in gene</th>
<th># of SNPs in cluster</th>
<th>Volume (mm³)</th>
<th>Proportion of brain volume</th>
<th>ClusterMax (mm³)</th>
<th># of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>GAB2</td>
<td>20</td>
<td></td>
<td>6336</td>
<td>0.0049</td>
<td>2688</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>LRDD</td>
<td>2</td>
<td></td>
<td>8128</td>
<td>0.0063</td>
<td>7872</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>PTPRB</td>
<td>17</td>
<td></td>
<td>3200</td>
<td>0.0024</td>
<td>3008</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>ZNF462</td>
<td>9</td>
<td></td>
<td>2688</td>
<td>0.0021</td>
<td>2688</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>IGSF5</td>
<td>27</td>
<td></td>
<td>16,384</td>
<td>0.013</td>
<td>9344</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>SLC25A12</td>
<td>10</td>
<td>5</td>
<td>1856</td>
<td>0.0014</td>
<td>1792</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>MRE11A</td>
<td>9</td>
<td>6</td>
<td>9344</td>
<td>0.0072</td>
<td>9216</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>SLC8A2</td>
<td>11</td>
<td>7</td>
<td>5632</td>
<td>0.0043</td>
<td>5376</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>CHRM5</td>
<td>3</td>
<td>3</td>
<td>1280</td>
<td>0.00099</td>
<td>1216</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>SPIRE1</td>
<td>19</td>
<td>14</td>
<td>6016</td>
<td>0.0046</td>
<td>3072</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>C3orf64</td>
<td>9</td>
<td>8</td>
<td>4352</td>
<td>0.0034</td>
<td>2112</td>
<td>4</td>
</tr>
<tr>
<td>21</td>
<td>S100B</td>
<td>1</td>
<td>1</td>
<td>9344</td>
<td>0.0072</td>
<td>6656</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>CRCT1</td>
<td>1</td>
<td>1</td>
<td>4096</td>
<td>0.0032</td>
<td>3456</td>
<td>4</td>
</tr>
<tr>
<td>19</td>
<td>ZNF626</td>
<td>6</td>
<td>5</td>
<td>2560</td>
<td>0.002</td>
<td>2112</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>ELK4</td>
<td>1</td>
<td>1</td>
<td>4032</td>
<td>0.0031</td>
<td>2688</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>RSF1</td>
<td>8</td>
<td>6</td>
<td>768</td>
<td>0.00059</td>
<td>768</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>WFDC11</td>
<td>2</td>
<td>2</td>
<td>1280</td>
<td>0.00099</td>
<td>512</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>SCML4</td>
<td>27</td>
<td>18</td>
<td>3328</td>
<td>0.0026</td>
<td>3328</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>ERP27</td>
<td>8</td>
<td>14</td>
<td>2624</td>
<td>0.002</td>
<td>2176</td>
<td>2</td>
</tr>
</tbody>
</table>

AD Risk Gene (Reiman et al., 2007)
A Heteroscedastic Linear Model

\[ y_i(v) = x_i^T \beta(v) + z_i(c)^T \gamma(c,v) + e_i(v) \quad \text{for} \quad i = 1, ..., n \]

where \( \beta(v) = (\beta_1(v), ..., \beta_K(v))^T \) is a \( K \times 1 \) vector associated with non-genetic predictors, and \( \gamma(c,v) = (\gamma_1(c,v), ..., \gamma_L(c,v))^T \) is an \( L \times 1 \) vector of genetic fixed effects (e.g., additive or dominant).

Moreover, \( e_i(v) \) are measurement errors with zero mean and \( e_i = \{ e_i(v) : v \in V \} \) are independent across \( i \).
We need to test:

\[ H_0(c,v) : \gamma(c,v) = 0 \] versus \[ H_1(c,v) : \gamma(c,v) \neq 0 \] for each \((c,v)\)

We calculate a Wald-type statistic as:

\[
W(c,v) = \tilde{\gamma}(c,v)^T \left\{ \text{Cov}(\tilde{\gamma}(c,v)) \right\}^{-1} \tilde{\gamma}(c,v)
\]

\[
= \text{tr} \left\{ \left\{ Z_c^T (I_n - P_X) Z_c \right\}^{-1} Z_c^T (I_n - P_X) \sigma_e^{-2}(c,v) Y(v) Y(v)^T (I_n - P_X) Z_c \right\}
\]
Big-Data Challenges

Memory:

\[ O((p_x + p_y)n + p_x p_y) \]

Computational time:

\[ O(p_x p_y n) = O(10^{17}) \]
Several big-data challenges arise from the calculation of $W(c,v)$ as follows.

- Calculating $\sigma^2_e(c,v)$ across all $(c,v)$'s can be computationally intensive.
- Holding all $W(c,v)$ in the computer hard drive requires substantial computer resources.
- Speeding up the calculation of $W(c,v)$.
FVGWAS

(I) Spatially Heteroscedastic Linear Model

(II) Global Sure Independence Screening Procedure

(III) Detection Procedure
Key Features

\[ X : p_x \times 1 \]

\[ B^T : p_y \times p_x \]

\[ Y \]

\[ \tilde{X} : p_x \times 1 \Rightarrow \tilde{X}^R : p_x^R \times 1 \]

\[ \tilde{B}^T : p_y \times p_x \]

\[ X: \text{Sparsity}; \quad Y|X: \text{Clustered ROIs} \]

UNIVERSITY of ALBERTA
To solve these computational bottlenecks, we propose two solutions as follows.

• Calculate $\sigma^2(c,v)$ under the null hypothesis $H_0(c,v)$ for each $v$ and $c$.

• Develop a GSIS procedure to eliminate many ‘noisy’ loci based on a global Wald-type statistic.

❖ By using these two solutions, we are able to reduce the computational complexity from $O(p_x p_y n)$ to $O((p_x + p_y)n^2)$. 
The global Wald-type statistic at locus $c$ is defined as

$$W(c) = N^{-1}_v \text{tr} \left\{ Z_c^T (I_n - P_X) Z_c \right\}^{-1} Z_c^T (I_n - P_X) \left\{ \sum_{v \in V} \hat{\sigma}_e(v)^{-2} Y(v)Y(v)^T \right\} (I_n - P_X) Z_c \right\}$$

is independent of $c$,

the complexity of computing $\{W(c)\}$ is $O\left((p_x + p_y)n^2\right)$

the complexity of computing $\{W(c, v)\}$ is $O(p_x p_y n)$

$$\frac{p_x p_y}{\left((p_x + p_y)n\right)} = \frac{10^{6+7}}{\left(10^7 + 10^6\right)10^4} = 10^2$$

$$= \frac{10^{6+7}}{\left(10^7 + 10^6\right)10^3} = 10^3$$
Simulation settings: the dark, gray, and white regions in the figure, respectively, represent background, brain region, and the effected ROI associated with the causal SNPs.

Fig. Simulation results for comparisons between FVGWAS and the Matrix eQTL in identifying significant voxel-SNP pairs.
Results

Our computational time

About 33,800 s

Manhattan Plot

- Observed
- Expected $-\log_{10}(\rho)$

Chromosome
Fig. ADNI whole-brain GWAS: (a) the density plot of $\rho$ and its approximation; (b) the density plot of $\lambda$ and its approximation
Table. RAVEN map GWAS: significant voxel-SNP pairs at the 0.5 significance level (left) and significant cluster-SNP pairs at the 0.5 significance level (right)

<table>
<thead>
<tr>
<th>SNP</th>
<th>Number of voxel-SNP pairs</th>
<th>SNP</th>
<th>Number of cluster-SNP pairs</th>
<th>Max cluster p-value of the max cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2075650</td>
<td>23</td>
<td>rs11815438</td>
<td>1</td>
<td>7906</td>
</tr>
<tr>
<td>(TOMM40)</td>
<td></td>
<td></td>
<td></td>
<td>0.11</td>
</tr>
<tr>
<td>rs9490103</td>
<td>4</td>
<td>rs2480271</td>
<td>1</td>
<td>7365</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.23</td>
</tr>
<tr>
<td>rs2244634</td>
<td>2</td>
<td>rs7001339</td>
<td>1</td>
<td>6864</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.45</td>
</tr>
</tbody>
</table>
ADNI Data Analysis

Fig. ADNI whole-brain GWAS: selected slices of $-\log_{10}(p)$ for significant clusters corresponding to a SNP (rs2480271).
Motivating Ex: DTI Fiber Tract Data

- Diffusion properties (e.g., FA, RA)
  \[ Y_i(s_j) = (y_{i,1}(s_j), \ldots, y_{i,m}(s_j))^T \]
  \( \{ s_1, \ldots, s_{n_G} \} \)

- Grids
- Covariates (e.g., age, gender, diagnostic) \( \chi_1, \ldots, \chi_n \)
Decomposition:

\[ y_{i,k}(s) = x_i^T B_k(s) + z_i(g)^T \beta_k(s,g) + \eta_{i,k}(s) + \varepsilon_{i,k}(s) \]

Coefficients \[ x_1, \ldots, x_n \]

Long-range Correlation \[ \eta_{i,k}(\bullet) \sim SP(0, \Sigma_\eta) \]

Short-range Correlation \[ \varepsilon_{i,k}(\bullet) \sim SP(0, \Sigma_\varepsilon), \]

Covariance operator:

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

\[ \sqrt{n}\{\text{vec}(\hat{B}(d) - B(d) - 0.5O(H^2)) : d \in D\} \xrightarrow{L} G(0, \Sigma_B(d, d')) \]

Zhu, Li, and Kong (2012). AOS
<table>
<thead>
<tr>
<th>SNP ID</th>
<th>CHR ID</th>
<th>$-\log_{10}(p\text{-value})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1384333'</td>
<td>10</td>
<td>7.3057</td>
</tr>
<tr>
<td>rs2805816'</td>
<td>9</td>
<td>6.6177</td>
</tr>
<tr>
<td>rs34608777'</td>
<td>8</td>
<td>6.2466</td>
</tr>
<tr>
<td>2:170144362:T_TTA</td>
<td>2</td>
<td>6.1259</td>
</tr>
<tr>
<td>rs11263882'</td>
<td>1</td>
<td>5.8854</td>
</tr>
<tr>
<td>rs78723271'</td>
<td>15</td>
<td>5.7848</td>
</tr>
<tr>
<td>rs11580558'</td>
<td>10</td>
<td>5.7236</td>
</tr>
<tr>
<td>rs10841375'</td>
<td>12</td>
<td>5.6652</td>
</tr>
<tr>
<td>rs11910115'</td>
<td>21</td>
<td>5.6525</td>
</tr>
<tr>
<td>rs28839626'</td>
<td>7</td>
<td>5.5871</td>
</tr>
</tbody>
</table>
A Software for FVGWAS

FVGWAS: Fast Voxelwise Genome Wide Association Analysis

Covariates  Data  SNP  Image index  Image size

N0  G  Output directory  RUN  Clear