Advances in Statistical Methods in Imaging Genetics

Thomas Nichols, PhD
Departments of Statistics & Warwick Manufacturing Group
University of Warwick

International Imaging Genetics Conference
17 January 2011
Outline

• Optimal Imaging Phenotype
• Validating VBM Cluster Size Inference
• SNP Combining Methods
• Invalidating Heritability in Twin Studies
• Sparse Multivariate Association
Optimal Imaging Phenotype

• Desperate need for power
• fMRI imaging genetics
  – Only few regions have BOLD signal
  – Can’t ROI’s help? (Reduce multiple testing)
Power: 1 Test

- **Power**
  - The probability of rejecting $H_0$ when $H_A$ is true
- Specify the false positive rate, $\alpha$
  - Sets the cut-off
- Specify effect magnitude ($\Delta$) and variance ($\sigma^2$)
  - Sets the alternative, with mean $\Delta/\sigma$
Power: 100,000 Tests?

• Avoid Multiple Testing Problem if possible
  – Typically study will use well-characterized paradigm
  – Expected region of response should be known
• But…
  – Variation in functional and structural anatomy
  – “Perfect” region never known
• Should we use focal ROI?
• Voxel-wise search in neighborhood?
• Over whole brain anyway?
Qualitative Power Exploration

• Simplified power setting
  – Not voxel-wise; instead largish (>1000 voxel) VOIs
  – Large VOIs: Assuming $\sigma_{\text{within}} << \sigma_{\text{between}}$
    • Hence different sized VOI’s will have similar variance
  – Large VOIs: Assuming independence between VOIs

• Consider impact of many vs. fewer VOI’s
  – Many VOIs
    • Better follows anatomy, possible shape of signal
    • Worse multiple testing correction
  – Fewer VOIs
    • Will dilute localized signal
    • Fewer tests to correct for
AAL & Derived ROI Atlases

Atlas 0 (AAL)
k = 116 regions
\( \alpha_{FWE} = 0.00043 \)
(surrogate for correlated voxel-wise search)

Atlas 1 (AAL symmetric)
k = 58 regions
\( \alpha_{FWE} = 0.00086 \)

Atlas 2
k = 28 regions
\( \alpha_{FWE} = 0.00179 \)

Atlas 3
k = 17 regions
\( \alpha_{FWE} = 0.00294 \)

Atlas 4 (Lobar AAL)
k = 6 regions
\( \alpha_{FWE} = 0.00833 \)

Atlas 5 (whole GM)
k = 1 region
\( \alpha_{FWE} = 0.05000 \)
L Amygdala

Atlas 0 (AAL)
k = 116 regions
$\alpha_{FWE} = 0.00043$
Signal
  # VOIs = 1
  Strength = 100%

Atlas 3
k = 17 regions
$\alpha_{FWE} = 0.00294$
Signal
  # VOIs = 1
  Strength = 4.9%

Atlas 1 (AAL symmetric)
k = 58 regions
$\alpha_{FWE} = 0.00086$
Signal
  # VOIs = 1
  Strength = 47%

Atlas 4 (Lobar AAL)
k = 6 regions
$\alpha_{FWE} = 0.00833$
Signal
  # VOIs = 1
  Strength = 0.6%

Atlas 2
k = 28 regions
$\alpha_{FWE} = 0.00179$
Signal
  # VOIs = 1
  Strength = 47%

Atlas 5 (whole GM)
k = 1 region
$\alpha_{FWE} = 0.05000$
Signal
  # VOIs = 1
  Strength 0.1%
Power: L Amygdala, True ROI

- True ROI best (of course)
- Rich ROI atlas (k=116) beats coarser atlases
  - Dilution more punishing than greater multiple testing
Power: L Amygdala, Shifted ROI

- True ROI best
- Wrong (unshifted) ROI next
- Rich ROI atlas still beats coarser atlases
Power: $\frac{1}{2}$ of Mid-Cingulate

- Whole Mid-Cing ROI best
- Again, huge (k=116) atlas next best
- But we’ve assumed RFX
  - No precision gain for large ROI’s, as shrinking $\sigma_{\text{WiN}}$ is no help
Power: $\frac{1}{2}$ of Mid-Cingulate: FFX

- Whole Mid-Cing ROI best
- Now Symmetric AAL atlas (k=58) best!
  - If $\sigma_{BTW}$ small, precision increase with large ROIs has impact
Power Exploration Conclusions

- Compared Range of Scales
  - Whole Brain, Lobar (k=6),..., AAL (k=116)

- Focal structures – Focal ROI’s best
- More extended signals, with heterogeneity
  - Rich atlas best
    - Dilution of signal worse than Bonferroni

- But whole-brain always less powerful than reduced volume
  - Suggests voxel-wise result preferred, constrained coarsely
Outline

• Optimal Imaging Phenotype
• Validating VBM Cluster Size Inference
• SNP Combining Methods
• Invalidating Heritability in Twin Studies
• Sparse Multivariate Association
Validating SPM Inference

• “False positives in imaging genetics”
  – Showed SPM valid—conservative even—for imaging genetic data (i.e. huge n)
  – But only considered voxel-wise inference
Inference On Images: Voxel-wise vs. Cluster-wise

- **Voxel-wise**
  - Reject Ho, point-by-point, by statistic magnitude

- **Cluster-wise**
  - Define contiguous blobs with arbitrary threshold $u_{clus}$
  - Reject Ho for each cluster larger than $k_\alpha$
Cluster Inference & Stationarity

• Cluster-wise preferred over voxel-wise
  – Generally more sensitive
  – Spatially-extended signals typical

• Problem w/ VBM
  – Standard cluster methods assume stationarity, constant smoothness
  – Assuming stationarity, false positive clusters will be found in extra-smooth regions
  – VBM noise very non-stationary

• Nonstationary cluster inference
  – Must un-warp nonstationarity
  – Reported but not implemented
    • Hayasaka et al, NeuroImage 22:676–687, 2004
  – Now available as SPM toolbox
    • http://fmri.wfubmc.edu/cms/software#NS
Validating VBM Cluster Size Inference

• Data
  – 181 MCI subjects from ADNI
  – T1 MRI’s, optimized VBM w/ SPM5
  – 700 “Null” SNPs
    • MAF > 5%, HWE OK at 0.05/700
    • Evenly spaced from chromosome 3
    • Ch3? A chromosome w/ no candidate genes
      – APOE, PSEN1, PSEN2 and SORL1 specifically

• Analysis
  – Additive model for single SNP
    • If fewer than 10 rare homozygotes, merge with heterozygotes
    • Plus Age, Gender
  – 700 SPM results – count % w/ any false positives
  – Also, 10 permutations
    • Just in case any true association
Null Rejection Rates: T-test

- Voxel-wise results match M-L (slightly conservative)
- 6mm smoothing way off
- 12 mm smoothing only OK for 0.001 cluster-forming threshold!

<table>
<thead>
<tr>
<th></th>
<th>6mm smoothing</th>
<th>Rejection Rates</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$u_c$</td>
<td>Observed(%)</td>
<td>Permutated(%)$^1$</td>
<td>Observed(%)</td>
</tr>
<tr>
<td><strong>t-tests</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>cluster-size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stationary</td>
<td>0.001</td>
<td>10.7</td>
<td>9.2±1.1</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>23.3</td>
<td>24.8±2.2</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>47.4</td>
<td>46.7±1.8</td>
<td>20.7</td>
</tr>
<tr>
<td>non-stationary</td>
<td>0.001</td>
<td>10.0</td>
<td>8.1±1.0</td>
<td>3.9</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>19.9</td>
<td>21.2±1.6</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>45.0</td>
<td>43.2±2.2</td>
<td>20.1</td>
</tr>
<tr>
<td><strong>voxel-wise</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FWE</td>
<td>-</td>
<td>3.4</td>
<td>2.7±0.6</td>
<td>3.0</td>
</tr>
<tr>
<td>FDR</td>
<td>-</td>
<td>3.3</td>
<td>2.2±0.6</td>
<td>1.9</td>
</tr>
</tbody>
</table>
Null Rejection Rates: F-test

- Similar pattern
  - Voxel-wise OK
  - Cluster-wise only OK for 12mm, 0.001!

<table>
<thead>
<tr>
<th>$f$-tests</th>
<th>$u_c$</th>
<th>6mm smoothing</th>
<th>12mm smoothing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Observed(%)</td>
<td>Permtued(%)$^1$</td>
</tr>
<tr>
<td>stationary</td>
<td>0.001</td>
<td>13.0</td>
<td>10.3±1.6</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>30.9</td>
<td>29.6±2.1</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>60.4</td>
<td>60.6±1.7</td>
</tr>
<tr>
<td>non-stationary</td>
<td>0.001</td>
<td>11.6</td>
<td>9.1±1.2</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>25.6</td>
<td>25.1±2.1</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>57.6</td>
<td>55.6±2.0</td>
</tr>
<tr>
<td>voxel-wise</td>
<td>FWE</td>
<td>-</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>-</td>
<td>2.9</td>
</tr>
</tbody>
</table>
(In)Validating VBM Cluster Size Inference

• Non-stationary cluster-size test OK in small samples!
  – Result of Hayasaka et al, in permutation and simulation

• What’s happening here?

• Check with simulations
  – Simulate same N=181
  – Stationary & non-stationary
    • FWHM 4,6,9 & 8,12,18
  – Count false positives out of 700 analyses
Null Rejection Rates: T-test Simulated Data

- Accurate to cluster-forming 0.01
  - Matching Hayasaka et al.

<table>
<thead>
<tr>
<th>t-tests</th>
<th>Rejection Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6mm smoothing</td>
</tr>
<tr>
<td></td>
<td>$u_c$</td>
</tr>
<tr>
<td>stationary</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>non-stationary</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>voxel-wise</td>
<td></td>
</tr>
<tr>
<td>FWE</td>
<td>-</td>
</tr>
<tr>
<td>FDR</td>
<td>-</td>
</tr>
</tbody>
</table>
Real data vs. Simulated Data
Why the mis-match?

• Simulated non-stationary data
  – Gaussian assumptions true
  – Form of non-stationary simple

• Real VBM data
  – Gaussianity possibly violated
    • But Central Limit Theorem should help
  – Complex pattern of nonstationarity
    • Theory depends on existence of a warp to stationarity

• More reason to use permutation!
Valid VBM Cluster-Size Inference

• Must account for nonstationarity
• For smooth data, cluster-forming threshold 0.001
  – RFT-based test OK
• Other-wise, permutation
  – Permutation valid under nonstationarity!
    • But not uniformly sensitive
  – Non-stationarity permutation test
    • Not currently implemented
      – SnPM & randomise will get it
Outline

- Optimal Imaging Phenotype
- Validating VBM Cluster Size Inference
- SNP Combining Methods
- Invalidating Heritability in Twin Studies
- Sparse Multivariate Association
Genetics Background

• SNPs vs. genes
  – Each gene often has several variants
  – 1 or more (but not many) SNPs typically needed to identify a gene
  – SNPs may not lie directly on coding portion of gene
    • Due to linkage disequilibrium (correlation), close is good enough
    • Non-coding, regulatory region may be causal
Modelling Multiple SNPs

• Model each SNP separately
  – But have yet another multiple testing problem!

• Haplotype
  – A ‘multivariate’ approach
  – Fit separate effect for each combination of SNPs

Haplotypes vs. Combined SNPs

• Statistical Geneticists Conventional Wisdom:
  – Haplotypes not worth the DF
  – Separate modelling of SNPs better

• Simulation study
  – Chapman et al., Human Heredity 2003; 56:18–31
  – Compared association power for several type 1 diabetes candidate genes
  – Induce effect at a single causal SNP
  – Vary degrees of freedom (DF) of model
    • Lowest DF: 1 covariate per SNP
    • Highest DF: All possible haplotypes
Haplotypes vs. Combined SNPs

• When few SNPs available...
  – Power doesn’t increase with haplotypes
Haplotypes vs. Combined SNPs

- When many SNPs available...
  - Simplest model almost always best
  - Effect is lost among a sea of DF

See also:
How to Combine Separate SNP Results?

- Tippett’s Method (1931)
  - Minimum P-value
  - Intuitive, corresponds to picking best result

- Fisher’s Method (1950)
  - Based on product of P-values
  - Equivalently, sum log P: $-2 \times \sum \log P_i \sim \chi^2_{2n}$

- Stouffer’s Method (1949)
  - Scaled Average Z, $\text{AvgZ} \times \sqrt{n} \sim N(0,1)$
  - If only have P, convert with $Z = \Phi(1-P)$

  - Max P-Value
Combining Methods: Interpretations

• Tippett’s Method – Min P
  – Can identify individual significant SNPs
  – Relatively better for rare/sparse signals

• Fisher’s Method – Prod P
  – Relatively better for diffuse/distributed signals
  – Detects regardless of inconsistent signs of effects

• Stouffer’s Method – Avg Z
  – Different signs of effects cancel

• Conjunctions – Max P
  – Relatively sensitive to consistent significance
Combining Methods: Other Variations

- **Weighted Stouffer’s Method**
  - $\sum w_i Z_i$ such that $\sum w_i^2 = 1$
  - Any weights; typically $w_i \propto \sqrt{n_i}$

- **Truncated Product Method**
  - Product of P-values smaller than limit $\lambda$

- **Rank Truncated Product Method**
  - Product of $k$ smallest P-values

- **Gamma Method**
  - Generalization of Fisher’s & Stouffer’s
  - Tuneable between small-but-consistent and large-but-rare signals
Combining Inference

• Inference on combining method
  – If combined P’s independent, can obtain parametric P-values for combined stat
    • e.g. \(-2 \times \sum_i \log P_i \sim \chi^2_{2n}\) for Fisher’s

• SNP’s not independent!
  – Need to use permutation, resampling methods
  – Under null, can permute SNP covariate
  – Must permute every SNP identically
  – Build permutation distribution of combining stat

• Implementation
  – randomise w/ fslmaths (FSL 4.1)
Permutation Inference for Combining Statistics

- Uncorrected P-values: C compared to permutation distribution at each voxel
- FWE Corrected P-values: C compared maximum distribution of C (over space)
- Dependence: Any correlations between SNPs accounted for
Combining Demonstration

• VBM data
  – 278 subjects, Major Depressive Disorder & Healthy Controls
• Effect of interest
  – Differences in GM-gene associations between the two groups
  – Additive or recessive effect tested with F stat
• Genes of interest, from WNT pathway
  – DVL2 (3 SNPs)
  – WNT3a (7 SNPs)
  – KRM1 (12 SNPs)
  – ZEB2 (25 SNPs)
• Combining over SNPs
  – Fishers, Stouffers, Max F & Min F
T images for 12 SNPs

Some T-images for interaction
Impossible to discern joint evidence by eye
Fisher’s & Stouffer’s best for weak but common effects.
  - No F larger than 5.
Max F best for rare, very strong effects.
- Very large F-value for 1st SNP.
Illustration with ZEB2 (3)

Fishers, Stouffer’s & Max F similar for common, strong effects.
Illustration with ZEB2 (4) 

Min F only best when no weak effects
- No large effects are required
Summary

• None attain FWE 0.05
  – ZEB2 close

• If any signal present, Max F most sensitive
Conclusions

• Imaging Genetics & Multiple SNPs
  – Except for well-characterised SNPs, generally can’t pick single SNP in advance
  – Must use principled method for combining

• Combining Methods
  – Right method depends on expected signal
    • Tippett’s: Best for rare, but intense effects
    • Fisher’s: Good for common, less intense effects
  – Expect point-mutations but linkage disequilibrium means SNPs may be correlated
Outline

• Optimal Imaging Phenotype
• Validating VBM Cluster Size Inference
• SNP Combining Methods
• Invalidating Heritability in Twin Studies
• Sparse Multivariate Association
Heritability from Twin Data

• Correlation between phenotype data of twins hold information about genetic heritability of traits
  – No correlation → No shared genetic variance
  – Correlation → Shared genetic and/or environmental variance

• Need both types of twins
  – Modelling both
    • monozygotic (MZ, identical) twins’ correlations, and
    • dizygotic (DZ, fraternal) twins’ correlation
      allows dissociation of genetic and environmental contributions

• Simple Falconer’s method uses arithmetic of correlation coefficients
  – Used extensively in imaging
    • E.g. Wright et al, NeuroImage 17, 256–271 (2002), 10 MZ, 10 DZ

• Better models use ‘SEM’
  – But here SEM is nothing more than window-dressing
  – Can work with components-of-variance model instead
\[ \text{Cov}(Y_1, Y_2) = \]

\[
\begin{bmatrix}
A + D + C + E & A + D + C \\
A + D + C & A + D + C + E
\end{bmatrix}
\]

\[
\begin{bmatrix}
A + D + C + E & \frac{1}{2}A + \frac{1}{4}D + C \\
\frac{1}{2}A + \frac{1}{4}D + C & A + D + C + E
\end{bmatrix}
\]

\[
\begin{bmatrix}
A + D + C + E & A + D \\
A + D & A + D + C + E
\end{bmatrix}
\]

\[
\begin{bmatrix}
A + D + C + E & \frac{1}{2}A + \frac{1}{4}D \\
\frac{1}{2}A + \frac{1}{4}D & A + D + C + E
\end{bmatrix}
\]

- “Narrow-sense” heritability
  - \( h^2 = A / (A + D + C + E) \)

- If all four types of twins are available, can estimate all four components
- If not, only really can estimate 3 parameters

\( A \) – Additive genetic variation
\( D \) – Dominant genetic variation
\( C \) – Common environmental variation
\( E \) – Random environmental / measurement variation
\[
\text{Cov}(Y_1, Y_2) = \begin{cases}
\text{From here on, assume raised together...}
\end{cases}
\]

\[
\begin{align*}
\text{MZ Twins} & \quad \text{No-Dominance model} & \quad \text{DZ Twins} \\
\begin{bmatrix}
A + C + E & A + C \\
A + C & A + C + E
\end{bmatrix} & \quad \begin{bmatrix}
A + C + E & \frac{1}{2} A + C \\
\frac{1}{2} A + C & A + C + E
\end{bmatrix}
\end{align*}
\]

\[
\begin{align*}
\text{MZ Twins} & \quad \text{No-common-environment model} & \quad \text{DZ Twins} \\
\begin{bmatrix}
A + D + E & A + D \\
A + D & A + D + E
\end{bmatrix} & \quad \begin{bmatrix}
A + D + E & \frac{1}{2} A + \frac{1}{4} D \\
\frac{1}{2} A + \frac{1}{4} D & A + D + E
\end{bmatrix}
\end{align*}
\]

\[
\begin{align*}
\text{MZ Twins} & \quad \text{Additive only, no-env model} & \quad \text{DZ Twins} \\
\begin{bmatrix}
A + E & A \\
A & A + E
\end{bmatrix} & \quad \begin{bmatrix}
A + E & \frac{1}{2} A \\
\frac{1}{2} A & A + E
\end{bmatrix}
\end{align*}
\]

\[
\begin{align*}
\text{MZ Twins} & \quad \text{No-genetic-effect model} & \quad \text{DZ Twins} \\
\begin{bmatrix}
C + E & C \\
C & C + E
\end{bmatrix} & \quad \begin{bmatrix}
C + E & C \\
C & C + E
\end{bmatrix}
\end{align*}
\]
Simulation Goals

• Use SPM ReML / Covariance modelling framework
  – i.e. $\text{Cov}(Y) = \sum \lambda_i Q_i$
  – One $\lambda$ for each of A, C & E parameter
  – For now, univariate simulation (same code for mass-univariate)

• Can we accurately pick the right model?
  – Simulate 4 different models as truth
  – For each given simulation, fit all 4 models, pick one with best LE

• Bias?
  – Not in the world of linear models, can’t assume unbiased

• Variance
  – Is our estimator better than dumb algebraic “Falconer’s estimate”
  – $h^2 = 2 (r_{MZ} - r_{DZ})$
Models Considered in Initial Simulation

<table>
<thead>
<tr>
<th>MZ Twins</th>
<th>DZ Twins</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Null” - No correlation model</td>
<td></td>
</tr>
<tr>
<td>$\begin{bmatrix} E &amp; 0 \ 0 &amp; E \end{bmatrix}$</td>
<td></td>
</tr>
<tr>
<td>“A only” Additive only, no-env model</td>
<td></td>
</tr>
<tr>
<td>$\begin{bmatrix} A + E &amp; A \ A &amp; A + E \end{bmatrix}$</td>
<td></td>
</tr>
<tr>
<td>“C only” - No-genetic-effect model</td>
<td></td>
</tr>
<tr>
<td>$\begin{bmatrix} C + E &amp; C \ C &amp; C + E \end{bmatrix}$</td>
<td></td>
</tr>
<tr>
<td>“A&amp;C” - No-Dominance model</td>
<td></td>
</tr>
<tr>
<td>$\begin{bmatrix} A + C + E &amp; A + C \ A + C &amp; A + C + E \end{bmatrix}$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MZ Twins</th>
<th>DZ Twins</th>
</tr>
</thead>
<tbody>
<tr>
<td>“A&amp;C” - No-Dominance model</td>
<td></td>
</tr>
<tr>
<td>$\begin{bmatrix} A + C + E &amp; \frac{1}{2} A + C \ \frac{1}{2} A + C &amp; A + C + E \end{bmatrix}$</td>
<td></td>
</tr>
</tbody>
</table>
Null
(no corr.)

A only
(no shared env.)

C only
(no heritability)

A & C
(shared genes & env.)
Bias Comparison: ReML $h^2$ (blue) vs. Falconer’s $h^2$ (red)

10 MZ + 10 DZ twins

30 MZ + 30 DZ twins

50 MZ + 50 DZ twins

“Null” (no corr.)
“A only” (no shared env.)
“C only” (no heritability)
“A & C” (shared genes & env.)
Stdev. Comparison: ReML $h^2$ (blue) vs. Falconer’s $h^2$ (red)

- 10 MZ + 10 DZ twins
- 30 MZ + 30 DZ twins
- 50 MZ + 50 DZ twins

“Null” (no corr.)  
“A only” (no shared env.)  
“C only” (no heritability)  
“A & C” (shared genes & env.)
MSE Comparison: ReML $h^2$ (blue) vs. Falconer’s $h^2$ (red)

- **50 MZ + 50 DZ twins**
  - Null
  - A only
  - C only
  - A & C

- **30 MZ + 30 DZ twins**
  - Null
  - A only
  - C only
  - A & C

- **10 MZ + 10 DZ twins**
  - Null
  - A only
  - C only
  - A & C

“Null” (no corr.)  “A only” (no shared env.)  “C only” (no heritability)  “A & C” (shared genes & env.)
Simulation Results

• Falconer’s method
  – Anticonservative
• REML
  – Much more accurate
• Bad news
  – Very hard to pick right model
  – Need lots of data
• Model Selection Difficulties
  – Known in the literature
    • Almost impossible to detect non-zero C
  – Recommended to always use A+C+E model
Outline

• Optimal Imaging Phenotype
• Validating VBM Cluster Size Inference
• SNP Combining Methods
• Invalidating Heritability in Twin Studies
• Sparse Multivariate Association
Conclusions

• Voxel-wise (or many smaller vs. fewer bigger ROIs) seem best
  – But of course with smallest mask possible

• Multiple SNPs often inescapable – don’t muddle through, combine!

• Statistical genetics has a rich literature – check there before grabbing a tool (or developing a new one.
Possible Mass-Univariate Analyses

- Full cross analysis
  - Massive multiple testing problem!

- Candidate SNP
  - Full image result
  - Must have right SNP

- Voxel/Region QTL
  - Whole genome association
  - Must have right ROI
True Joint Imaging-Genetics Modelling

- Ideally want to fit all SNPs, all voxels/ROIs
- Need multivariate method
  - To eliminate avoid multiple testing
- Need sparsity
  - To implicitly do testing

Joint work with Maria Vounou, Giovanni Montana, Imperial College
Multivariate Regression

Image Data \[ Y \]
\[ n \times q \]

SNP Data \[ X \]
\[ n \times p \]

Regression Coefficients \[ C \]
\[ p \times q \]

Error \[ E \]
\[ n \times q \]

• Impossible to fit
  – unless \( n \) exceeds \( p \) or \( q \)
Multivariate Reduced Rank Regression

\[ Y = X \begin{pmatrix} A & B \end{pmatrix} + E \]

- Feasible to estimate
  - But not easy to interpret
Multivariate Sparse Reduced Rank Regression

- Mostly-zero coefficients
  - Combine model-fitting and selection

\[ Y = X + A B + E \]

- Image Data
- SNP Data
- Sparse Imaging Coefficients
- Error
Reduced Rank Regression Analysis (1)

- **Standard RRR**
  - \( X \): SNP data \((n_{\text{subj}} \times n_{\text{SNP}})\)
  - \( Y \): Imaging data \((n_{\text{subj}} \times n_{\text{voxel}})\)
  - For given rank \( r \), find \( B \times A \) that approximates true \( C \)

- **Sparse RRR**
  - Iteratively find vectors \( a \) & \( b \) to minimize
    \[
    \text{tr}\{ (Y - X ba) (Y - X ba)' \} + \lambda_a ||a'||_1 + \lambda_b ||b||_1
    \]

- **What penalty?**
  - \( L_1 \) “Lasso”
  - Forces some elements exactly zero
Evaluating Sparse RRR for SNP-MRI Association

• Structural MRI data
  – ADNI T1 images through SPM5 VBM pipeline
  – ~100 AD subjects
  – Reduce dimensionality with ROI template
    • 10 coarse ROIs
      – Occipital, Parietal, Temporal, Frontal, Insular, Cingulate, Thalamus, ...
    • Estimate covariance matrix V after adjusting for age & gender

• Evaluate with realistic genetic population w/ FREGENE
  – Simulates sequence-level data in large population
  – Provides 10K individuals, 20Mb chromosome (~180K SNPs)
FREGENE simulation example
World population using

Why try so hard? Why not just rand\{0,1,2\}^{500,000}?

- Linkage disequilibrium (LD)
  - SNPs not independent
  - Highly structured, heterogeneous dependence
- Population sub-structure
  - Ethnic differences & migration patterns induce systematic variation
- Want confidence that our multivariate method performs well with such structure
Evaluating Sparse RRR for SNP-MRI Association

• Simulated SNP data
  – Use 300 SNPs from 5 Kb region
  – From population of 10,000, repeatedly sample 2,000 cohorts
  – Selected 4 *causative* SNPs
    • Will be used to induce phenotypic effect
    • But then *dropped* from set, leaving 296 SNPs
    • Represents realistic setting, where causative SNP is not seen, but effect captured through local LD

• Simulated MRI data
  – Simulate ROI data with covariance V
  – Add genetic effect to Frontal and Temporal ROIs with causative SNPs

• “True positive” with missing causative SNP
  – Declare true positive if LD coefficient close enough
ROC Components – SNP Components

For $r=1$, same or better than Mass Univariate
For $r>1$, always better
ROC Performance – SNP Components: More SNPs

Generalization to 10’s of 1,000’s of SNPs… always out-performs Mass Univariate
ROC Performance – Imaging Components

For equal weighting, not ROI as great
Unequal weighting somewhat better
Future Work
Whole Brain / Whole Genome Association

• Computationally hard, not impossible
  – Weeks, not months of computing
  – Various tricks available
    • Parametric – Use fast, approximate method to screen out large P’s
    • Nonparametric - Use PLINK-like tricks

• Must contemplate power issues
  – Imaging power per sample surely higher than case/control outcome...
  – *But*, imaging sample sizes rarely approach those of case/control studies
  – Must be able to argue that power exists to find associations that survive the {brain} x {genome} search

• Genetic Modelling
  – Single locus method?
  – Running multi-locus method?
Conclusions

• Imaging Genetics
  – Mash-up of two large data, massive multiple testing problems

• Candidate SNP VBM
  – Need narrow primary outcome definition
  – But secondary outcomes, duly qualified, valuable for generating hypotheses

• Need more work on SNP selection, joint modelling/inference
Acknowledgements

• GlaxoSmithKline
  – Becky Inkster
    • Data QC, VBM Model fitting, gene selection,
  – Anil Rao, Brandon Whitcher
    • VBM Preprocessing
  – Peirandrea Muglia, Paul Matthews
    • Project management, data curation, gene selection

• External
  – Satoru Hayasaka, Wake Forest University
    • Nonstationary Cluster Inference Toolbox