Analyzing matrix data through Bayesian multilevel modeling

Introduction to the AFNI program MBA

Gang Chen

SSCC/NIMH/NIH

March 11, 2021

Overview

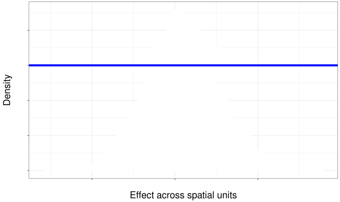
- Conventional approach: massively univariate analysis
  - Advantages and limitations
- New framework: Bayesian multilevel (BML) modeling
  - Model formulation
  - Overcoming the limitations of MUA
  - AFNI program MBA
- Demo: running MBA with matrix data

Staple Methodology in Neuroimaging

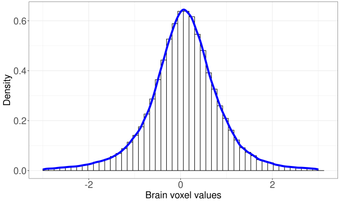
- Massively univariate analysis (MUA)
  - Spatial units: voxels, ROIs, region pairs
  - Treat each spatial unit as an isolated entity
  - As many models as spatial units: GLM, ANOVA, LME, NBS
  - No commonality across peers: trees/patches, but no forest!
  - Intuitive and straightforward: flat, usually not hierarchical
- Good and bad
  - + Simple and flat: GLM at the unit level; Divide and Conquer
  - + Computationally frugal
  - – Heavy penalty: multiple testing adjustment (MTA)
    - Family-wise error (FWE): cluster-based, permutation-based
    - False discovery rate (FDR)
  - – Artificial dichotomy: draw an arbitrary line in sand (how "correct" is 0.05?)
  - – Information waste: MTA and dichotomization intertwined

Problem with MUA

- No relationship across space



- Really no prior knowledge?



- Damages
  - Overfitting: poor predictivity / reproducibility
  - Information waste
  - Heavy penalty
  - Dichotomization
  - Discrimination against anatomically small regions
  - Vulnerability / sensitivity to data manipulations

Matrix Data

- Data types
  - FMRI: inter-region correlations
  - DTI: white-matter properties: FA, MD, RD, AD, ...

- Structure
  - m ROIs $(i, j = 1, 2, .., \ m(i \neq j))$, n subjects $(k = 1, 2, ..., n)$

$$
\boldsymbol{Z}_k^{(m)} = \begin{array}{c} \\ R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_m \end{array}
\begin{array}{c}
\begin{array}{ccccc} R_1 & R_2 & R_3 & \cdots & R_m \end{array} \\
\begin{pmatrix}
- & z_{12k} & z_{13k} & \cdots & z_{1mk} \\
z_{21k} & - & z_{23k} & \cdots & z_{2mk} \\
z_{31k} & z_{32k} & - & \cdots & z_{3mk} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
z_{m1k} & z_{m2k} & z_{m3k} & \cdots & -
\end{pmatrix}
\end{array}
$$

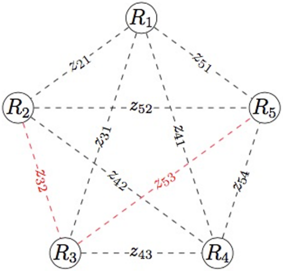
  - Usually symmetric: $\frac{1}{2}m(m-1)$ elements z_{ijk}
  - Missing elements?

Matrix Data: MUA

- GLM
  - $\frac{1}{2}m(m-1)$ models: one per element
  - Implicit assumptions
    - All region pairs: unrelated; no information shared
    - Equal likelihood of all ROIs being in $(-\infty, +\infty)$
    - Divide and Conquer: see trees, but not forest
  - Tools: NBS, CONN, FSLNets in FSL, GIFT
- Costs
  - Heavy penalty: multiplicity
  - Binarization
  - information waste, poor reproducibility

Examining implicit assumptions: MUA

- Unrelatedness among region pairs?



$$\boldsymbol{P}^{(5)} = \begin{array}{c} \\ z_{21} \\ z_{31} \\ z_{41} \\ z_{51} \\ z_{32} \\ z_{42} \\ z_{52} \\ z_{43} \\ z_{53} \\ z_{54} \end{array} \begin{array}{c} \begin{array}{cccccccccc} z_{21} & z_{31} & z_{41} & z_{51} & z_{32} & z_{42} & z_{52} & z_{43} & z_{53} & z_{54} \end{array} \\ \begin{pmatrix} 1 & \rho & \rho & \rho & \rho & \rho & \rho & 0 & 0 & 0 \\ \rho & 1 & \rho & \rho & \rho & 0 & 0 & \rho & \rho & 0 \\ \rho & \rho & 1 & \rho & 0 & \rho & 0 & \rho & 0 & \rho \\ \rho & \rho & \rho & 1 & 0 & 0 & \rho & 0 & \rho & \rho \\ \rho & \rho & 0 & 0 & 1 & \rho & \rho & \rho & \rho & 0 \\ \rho & 0 & \rho & 0 & \rho & 1 & \rho & \rho & 0 & \rho \\ \rho & 0 & 0 & \rho & \rho & \rho & 1 & 0 & \rho & \rho \\ 0 & \rho & \rho & 0 & \rho & \rho & 0 & 1 & \rho & \rho \\ 0 & \rho & 0 & \rho & \rho & 0 & \rho & \rho & 1 & \rho \\ 0 & 0 & \rho & \rho & 0 & \rho & \rho & \rho & \rho & 1 \end{pmatrix} \end{array}$$

  - How to capture the relatedness ρ? Not addressed under MUA

- Equal likelihood of all ROIs being in $(-\infty, +\infty)$?
  - ROIs: similar to subjects and trials in task-related FMRI
  - Gaussian distribution: more reasonable than uniform

Integration through multilevel modeling

- Data: m ROIs $(i, j\ (i \neq j))$, n subjects (k)

$$\boldsymbol{Z}_k^{(m)} = \begin{array}{c} \\ R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_m \end{array} \begin{array}{c} \begin{array}{ccccc} R_1 & R_2 & R_3 & \cdots & R_m \end{array} \\ \begin{pmatrix} - & z_{12k} & z_{13k} & \cdots & z_{1mk} \\ z_{21k} & - & z_{23k} & \cdots & z_{2mk} \\ z_{31k} & z_{32k} & - & \cdots & z_{3mk} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{m1k} & z_{m2k} & z_{m3k} & \cdots & - \end{pmatrix} \end{array}$$

- Model formulation
  - $z_{ijk} \sim \mathcal{N}(\mu, \sigma^2)$, $\mu = b_0 + \xi_i + \xi_j + \eta_{ij} + \zeta_{ik} + \zeta_{jk} + \pi_k$; $\xi_i, \xi_j \sim \mathcal{N}(0,\ \lambda^2)$, $\eta_{ij} \sim \mathcal{N}(0,\ \theta^2)$, $\zeta_{ik}, \zeta_{jk} \sim \mathcal{N}(0,\ \nu^2)$, $\pi_k \sim \mathcal{N}(0,\ \tau^2)$; $i, j = 1, 2, ..., m$; $k = 1, 2, ..., n$
  - ROIs loosely constrained/regularized: $\xi_i, \xi_j \sim \mathcal{N}(0,\ \lambda^2)$

Multilevel modeling: making inferences

- Conceptual model
  - Effect decomposition $\mu = b_0 + \xi_i + \xi_j + \eta_{ij} + \zeta_{ik} + \zeta_{jk} + \pi_k$
  - Effects of interest
    - Region pair: $b_0 + \xi_i + \xi_j + \eta_{ij}$
    - Region: $\frac{1}{2}b_0 + \xi_i$
  - Relatedness among region pairs characterized: ρ
  - Integrity: see the forest for the trees
  - No multiplicity/dichotomization: one model integrates all
- Numerical implementations
  - Linear mixed-effects (LME) modeling
    - Computationally scalable, but algorithmically infeasible
  - Bayesian multilevel (BML) framework: probabilistic modeling
    - Algorithmically easy with manageable computational cost
    - Inferences via simulations: Markov chain Monte Carlo (MCMC)
    - Implemented in AFNI program MBA

Demo: matrix-based analysis with MBA

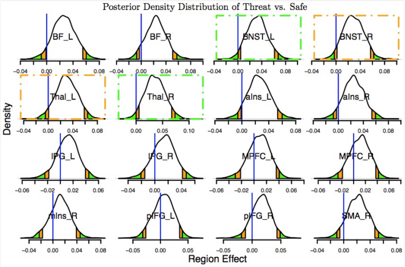
- Dataset: correlation matrices
  - Response-conflict task (Choi et al., 2012)
  - $n = 41$ subjects, $m = 16$ ROIs; 120 correlations per subject
- Input: data.txt in long format

```
Subj  ROI1  ROI2     Y
 S1    R1    R2  -0.1625
 S1    R1    R3   0.0238
 S1    R1    R4   0.1846
 ...
```

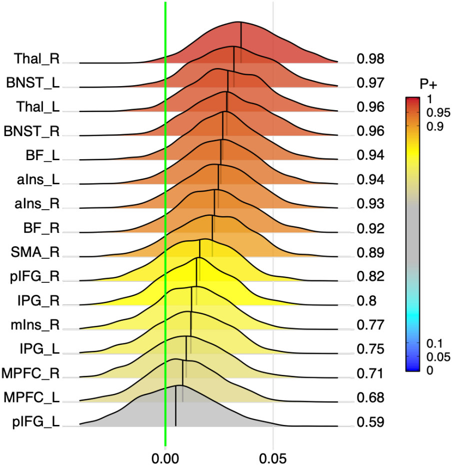
- MBA script
  MBA -prefix output -chains 4 -WCP 6 -iterations 1000 -EOI 'Intercept' -dataTable data.txt

MBA: Inferences at ROI level

- Distribution for each ROI
- Relative importance: hub
- Full result presentation
- Highlight but not hide

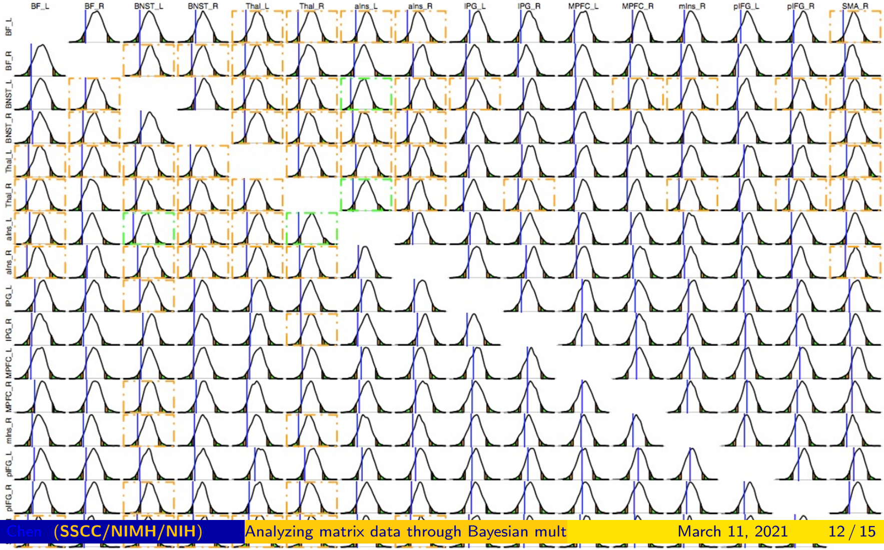


Ridge plot



MBA: Inferences at region pair level

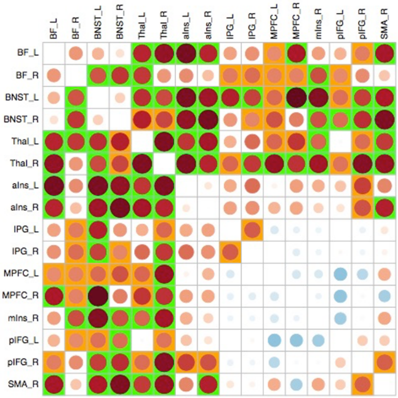
- 120 region pairs



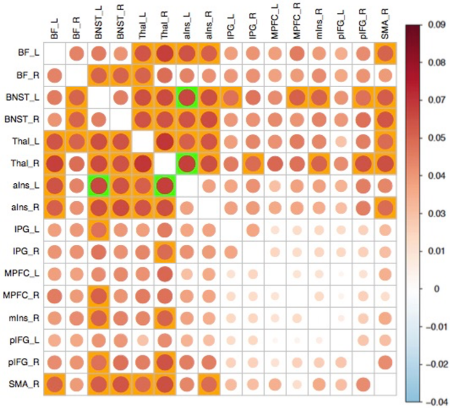
Inferences at region pair level

- MUA: 63 RPs with $p = 0.05$, but 0 under MTA via NBS
- MBA: 33 RPs with strong evidence under BML

MUA/GLM



MBA/BML



MBA applied to DTI data

- Model validation?
- Limited applications
- Challenge: missing region pairs
  - Modeling: trouble with too many missing RPs
  - Result presentation

Acknowledgements

- Paul Taylor, Bob Cox, Rick Reynolds, Daniel Glen, John Lee, Justin Rajendra,
  SSCC, NIMH/NIH
- Yaqiong Xiao, Elizabeth Redcay, Tracy Riggins, Fengji Geng, Luiz Pessoa, Joshua Kinnison,
  Depart of Psychology, University of Maryland