Multivariate Modeling (MVM): A Comprehensive Approach to Group Analysis

Gang Chen, Ph.D.

Scientific and Statistical Computational Core

National Institute of Mental Health

National Institutes of Health, USA

http://afni.nimh.nih.gov/sscc/gangc

Poster number 3606

Standby time: 12:45 – 14:45

Wednesday June 11

Also display time: Thursday

Group Analysis in NeuroImaging: why big models?

- ✧ Various group analysis approaches
  - Student's ***t*-test**: one-, two-sample, and paired
  - **ANOVA**: one or more categorical explanatory variables (factors)
  - **GLM**: AN(C)OVA
  - **LME**: linear mixed-effects modeling
- ✧ *t*-tests not always practical or feasible
  - Too tedious when layout is too complex
  - Main effects and interactions: desirable
  - When quantitative covariates are involved
- ✧ Advantages of big models: AN(C)OVA, GLM, LME
  - All tests in one analysis (vs. piecemeal *t*-tests)
  - Omnibus *F*-statistics
  - Power gain: combining subjects across groups

Piecemeal *t*-tests: 2 × 3 Mixed ANCOVA

- Explanatory variables
  - Factor A (Group): 2 levels (patient and control)
  - Factor B (Condition): 3 levels (pos, neg, neu)
  - Factor S (Subject): 15 ASD children and 15 healthy controls
  - Quantitative covariate: Age
- Multiple *t*-tests
  - Group comparison + age effect
  - Pairwise comparisons among three conditions
  - Effects that cannot be analyzed
    - Main effect of Condition
    - Interaction between Group and Condition
    - Age effect across three conditions

Classical ANOVA: 2 × 3 Mixed ANCOVA

- Factor A (Group): 2 levels (patient and control)
- Factor B (Condition): 3 levels (pos, neg, neu)
- Factor S (Subject): 15 ASD children and 15 healthy controls
- Quantitative covariate (Age): cannot be modeled with ANOVA

$$F_{(a-1,a(n-1))}(A) = \frac{MSA}{MSS(A)},$$

$$F_{(b-1,a(b-1)(n-1))}(B) = \frac{MSB}{MSE},$$

$$F_{((a-1)(b-1),a(b-1)(n-1))}(AB) = \frac{MSAB}{MSE},$$

where

$$MSA = \frac{SSA}{a-1} = \frac{1}{a-1}\left(\frac{1}{bn}\sum_{j=1}^{a} Y_{.j.}^2 - \frac{1}{abn}Y_{...}^2\right),$$

$$MSB = \frac{SSB}{b-1} = \frac{1}{b-1}\left(\frac{1}{an}\sum_{k=1}^{b} Y_{..k}^2 - \frac{1}{abn}Y_{...}^2\right),$$

$$MSAB = \frac{SSAB}{(a-1)(b-1)} = \frac{1}{(a-1)(b-1)}\left(\frac{1}{n}\sum_{j=1}^{a}\sum_{k=1}^{b} Y_{.jk} - \frac{1}{bn}\sum_{j=1}^{a} Y_{.j.}^2 - \frac{1}{an}\sum_{k=1}^{b} Y_{..k}^2 + \frac{1}{abn}Y_{...}^2\right),$$

$$MSS(A) = \frac{SSS(A)}{a(n-1)} = \frac{1}{a(n-1)}\left(\frac{1}{b}\sum_{i=1}^{n}\sum_{j=1}^{a} Y_{ij.}^2 - \frac{1}{bn}\sum_{j=1}^{a} Y_{.j.}^2\right),$$

$$MSE = \frac{1}{a(b-1)(n-1)}\left(\sum_{i=1}^{n}\sum_{j=1}^{a}\sum_{k=1}^{b} Y_{ijk}^2 - \frac{1}{n}\sum_{j=1}^{a}\sum_{k=1}^{b} Y_{.jk} - \frac{1}{b}\sum_{i=1}^{n}\sum_{j=1}^{a} Y_{ij.}^2 + \frac{1}{bn}\sum_{j=1}^{a} Y_{.j.}^2 + \frac{1}{abn}Y_{...}^2\right)$$

Univariate GLM: 2 x 3 mixed ANOVA

- Group: 2 levels (patient and control)
- Condition: 3 levels (pos, neg, neu)
- Subject: 3 ASD children and 3 healthy controls

Difficult to incorporate covariates

Subj
1
1
1
2
2
2
3
3
3
4
4
4
5
5
5
6
6
6

$$
\underbrace{\begin{pmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{13} \\ \beta_{21} \\ \beta_{22} \\ \beta_{23} \\ \beta_{31} \\ \beta_{32} \\ \beta_{33} \\ \beta_{41} \\ \beta_{42} \\ \beta_{43} \\ \beta_{51} \\ \beta_{52} \\ \beta_{53} \\ \beta_{61} \\ \beta_{62} \\ \beta_{63} \end{pmatrix}}_{b}
=
\underbrace{\begin{pmatrix}
1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 1 & -1 & -1 & -1 & -1 & 0 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 & 1 & 0 & -1 & -1 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & -1 & -1 & 0 & 0 \\
1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 \\
1 & -1 & 1 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\
1 & -1 & 0 & 1 & 0 & -1 & 0 & 0 & 1 & 0 \\
1 & -1 & -1 & -1 & 1 & 1 & 0 & 0 & 1 & 0 \\
1 & -1 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 1 \\
1 & -1 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 1 \\
1 & -1 & -1 & -1 & 1 & 1 & 0 & 0 & 0 & 1 \\
1 & -1 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & -1 \\
1 & -1 & 0 & 1 & 0 & -1 & 0 & 0 & -1 & -1 \\
1 & -1 & -1 & -1 & 1 & 1 & 0 & 0 & -1 & -1
\end{pmatrix}}_{X:\ X_0\ X_1\ X_2\ X_3\ X_4\ X_5\ X_6\ X_7\ X_8\ X_9}
\underbrace{\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \\ \alpha_8 \\ \alpha_9 \end{pmatrix}}_{a}
+
\underbrace{\begin{pmatrix} \delta_{11} \\ \delta_{12} \\ \delta_{13} \\ \delta_{21} \\ \delta_{22} \\ \delta_{23} \\ \delta_{31} \\ \delta_{32} \\ \delta_{33} \\ \delta_{41} \\ \delta_{42} \\ \delta_{43} \\ \delta_{51} \\ \delta_{52} \\ \delta_{53} \\ \delta_{61} \\ \delta_{62} \\ \delta_{63} \end{pmatrix}}_{d}
$$

Our Approach: Multivariate GLM

- Group: 2 levels (patient and control)
- Condition: 3 levels (pos, neg, neu)
- Subject: 3 ASD children and 3 healthy controls
- Age: quantitative covariate

$$B_{n\times m} = X_{n\times q}\, A_{q\times m} + D_{n\times m}$$

$$\begin{array}{c} Subj \\ 1\\2\\3\\4\\5\\6 \end{array}
\begin{array}{c} \begin{array}{ccc} Pos & Neg & Neu \end{array} \\ \underbrace{\begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \\ \beta_{31} & \beta_{32} & \beta_{33} \\ \beta_{41} & \beta_{42} & \beta_{43} \\ \beta_{51} & \beta_{52} & \beta_{53} \\ \beta_{61} & \beta_{62} & \beta_{63} \end{pmatrix}}_{B} \end{array}
=
\begin{array}{c} \begin{array}{ccc} Int & Grp & Age \end{array} \\ \underbrace{\begin{pmatrix} 1 & 1 & -6 \\ 1 & 1 & 10 \\ 1 & 1 & 4 \\ 1 & -1 & -4 \\ 1 & -1 & -1 \\ 1 & -1 & -3 \end{pmatrix}}_{X} \end{array}
\begin{array}{c} \begin{array}{ccc} Pos & Neg & Neu \end{array} \\ \underbrace{\begin{pmatrix} \alpha_{01} & \alpha_{02} & \alpha_{03} \\ \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \end{pmatrix}}_{A} \end{array}
+
\begin{array}{c} \begin{array}{ccc} Pos & Neg & Neu \end{array} \\ \underbrace{\begin{pmatrix} \delta_{11} & \delta_{12} & \delta_{13} \\ \delta_{21} & \delta_{22} & \delta_{23} \\ \delta_{31} & \delta_{32} & \delta_{33} \\ \delta_{41} & \delta_{42} & \delta_{43} \\ \delta_{51} & \delta_{52} & \delta_{53} \\ \delta_{61} & \delta_{62} & \delta_{63} \end{pmatrix}}_{D} \end{array}
\begin{array}{c} Subj \\ 1\\2\\3\\4\\5\\6 \end{array}$$

Univariate GLM: popular in neuroimaging

- Advantages: more *flexible* than the method of sums of squares
  - No limit on the the number of explanatory variables (in principle)
  - Easy to handle unbalanced designs
  - Covariates can be modeled when no within-subject factors present
- Disadvantages: costs paid for the flexibility
  - Intricate dummy coding
  - Tedious pairing for numerator and denominator of *F*-stat
    - Proper denominator SS
    - Can't generalize (in practice) to any number of explanatory variables
    - Susceptible to invalid formulations and problematic post hoc tests
  - Cannot handle covariates in the presence of within-subject factors
  - **No** direct approach to correcting for sphericity violation
    - Unrealistic assumption: **same** variance-covariance structure
- **Problematic**: When residual SS is adopted for all tests
  - *F*-stat: valid only for highest order interaction of within-subject factors
  - Most post hoc tests are inappropriate

Group Analysis: when GLM is not enough?

- ✧ Example: 5 factors + 1 covariate
  - 3 between-subjects factors
    - Group: adult, child; Diagnosis: healthy, anxious; Scanner: scanners 1 and 2
  - 2 within-subject factors: 3 × 3 at the individual level
    - Stimulus category: human, animal, tool; Emotion: pos, neg, neu
  - 1 quantitative covariate: Age
  - > 200 post-hoc tests + *F*-stats for main effects and interactions
  - Piecemeal *t*-test approach would not work
- ✧ **Three difficulties**: most packages cannot properly handle
  - Number of explanatory variables (factors and covariates): 6
  - Covariates in the presence of within-subject factors
  - Sphericity violation when > 2 levels for a within-subject factor
    - No direct method available under GLM
    - **Presumption**: same variance-covariance structure across the brain

Multivariate GLM for Univariate GLM / AN(C)OVA

✧ Classical multivariate testing: MAN(C)OVA

- Centroid testing for a within-subject factor with *m* levels
  - One-sample H_0: $(a_{pos}, a_{neg}, a_{neu}) = (0, 0, 0)$
  - Two-sample H_0: $(a_{1pos}, a_{1neg}, a_{1neu}) = (a_{2pos}, a_{2neg}, a_{2neu})$
- Usually not of interest for neuroimaging group analysis; instead
  - Main effect H_0: $a_{pos}=a_{neg}=a_{neu}$
  - Interaction H_0: $a_{1pos}-a_{2pos}=a_{1neg}-a_{2neg}=a_{1neu}-a_{2neu}$

✧ Hypothesis formulation H_0: $L_{u\times q}\, A_{q\times m}\, R_{m\times v} = C_{u\times v}$

- $L_{u\times q}$: weights for BS variables (groups and covariates)
- $R_{m\times v}$: weights for WS factor levels
- Example: 2 x 3 mixed ANOVA
- Construct statistics based on Sum of Squares and Products (SSP) matrices ***H*** and ***E*** for Hypothesis (SSPH) and Errors (SSPE)

Main Effect of A - $\boldsymbol{L}_A = (0,1,0), \boldsymbol{R}_A = (1,1,1)^T$

Main Effect of B - $\boldsymbol{L}_B = (1,0,0), \boldsymbol{R}_B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}$

Interaction A:B - $\boldsymbol{L}_{A:B} = (0,1,0), \boldsymbol{R}_{A:B} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}$

Multivariate GLM for Univariate Testing

- ✧ Univariate testing (UVT) for AN(C)OVA under MVM
  - F: $tr[\boldsymbol{H}(\boldsymbol{R}^T\boldsymbol{R})^{-1}] \;/\; tr[\boldsymbol{E}(\boldsymbol{R}^T\boldsymbol{R})^{-1}]$ scaled by DFs
- ✧ Bonuses in terms of modeling capability
  - No limit on the number of factors and covariates
  - Covariates can be modeled in presence of within-subject factors
  - Pairing for numerator and denominator of *F*-stats is automatic
  - Classical methods of correction for sphericity violations: Greenhouse-Geisser (GG) and Huynh-Feld (HF)
  - Convenient to perform post hoc tests
  - Multiple estimates of an effect (e.g., runs) handled automatically
  - Extra bonus: within-subject multivariate testing complementary to traditional UVT when sphericity violation is severe

Multivariate Testing under MVM

- Any effect involving a within-subject factor converted to a multivariate hypothesis: 2 x 3 mixed ANOVA
  - Main effect - B - H_0: $a_{pos}=a_{neg}=a_{neu}$ ⟺ H_0: $a_{pos}-a_{neu}=0$, $a_{neg}-a_{neu}=0$
  - Interaction H_0: $a_{1pos}-a_{2pos}=a_{1neg}-a_{2neg}=a_{1neu}-a_{2neu}$ ⟺
    H_0: $(a_{1pos}-a_{1neu}, a_{1neg}-a_{1neu})=(a_{2pos}-a_{2neu}, a_{2neg}-a_{2neu})$
- When HDR estimated with multiple basis functions
  - Univariate testing by reduction to scalar
    - Area under the curve (AUC)
    - Principal component
    - Summarized measure (Calhoun et al., 2004)
  - Comprehensive approach under MVM
    - AUC, main effect, interaction, MVT
- Other cases: multiple functional connectivity networks, multi-modality data analysis

MVM Implementation in AFNI

- ✧ Program 3dMVM
  - Command line
  - **Symbolic coding** for variables and post hoc testing

Variable types

Post hoc tests

```
3dMVM  -prefix       OutputFile      -jobs 8       -SC                          \
       -bsVars       'Grp*Age'       -wsVars       'Cond'    -qVars 'Age'       \
       -num_glt 4                                                               \
       -gltLabel 1   Pat_Pos         -gltCode 1           'Grp : 1*Pat Cond : 1*Pos'  \
       -gltLabel 2   Ctl_Pos-Neg     -gltCode 2     'Grp : 1*Ctl Cond : 1*Pos -1*Neg' \
       -gltLabel 3   GrpD_Pos-Neg    -gltCode 3 'Grp : 1*Ctl -1*Pat Cond : 1*Pos -1*Neg' \
       -gltLabel 4   Pat_Age         -gltCode 4                 'Grp : 1*Pat Age :'  \
       -dataTable                                                               \
       Subj          Grp             Age           Cond      InputFile          \
       S1            Ctl             23            Pos       S1_Pos.nii         \
       S1            Ctl             23            Neg       S1_Neg.nii         \
       S1            Ctl             23            Neu       S1_Neu.nii         \
       ...
       S50           Pat             19            Pos       S50_Pos.nii        \
       S50           Pat             19            Neg       S50_Neg.nii        \
       S50           Pat             19            Neu       S50_Neu.nii
```

Data layout

Summary

- Advantages of MVM
  - No limit on the number of explanatory variables
  - Covariates modeled even in the presence of within-subject factors
  - Voxel-wise covariate (e.g., SFNR) allowed
  - Voxel-wise sphericity correction for UVT
  - Easy and automatic formulation of testing statistics
  - Within-subject MVT as complementary testing
  - MVT: HDR modeled with multiple basis functions
- The user only provides information
  - Explanatory variable types: between- / within-subject, covariate
  - Centering options for quantitative covariates
  - Post hoc tests via symbolic coding
  - Data table listing variables and input files
- The user does not need to be involved in specifying
  - regressors, design matrix, and post hoc tests via regressors

Lastly

✧ Acknowledgements

- Robert C. Cox, Ph.D.
- Ziad S. Saad, Ph.D.
- Nancy E. Adleman, Ph.D.
- Ellen Leibenluft, M.D.
- NIMH+NINDS Intramural Research Programs, NIH/HHS/USA
- Statistical computational language and environment R

✧ More information

- **Poster** number 3606:
  - Standby time: 12:45 – 14:45 Wednesday June 11
  - Also display time: Thursday, June 12
- Website: http://afni.nimh.nih.gov/sscc/gangc
- Paper: Chen et al., Applications of Multivariate Modeling to Neuroimaging Group Analysis: A Comprehensive Alternative to Univariate General Linear Model, *NeuroImage* (reviewer 1 permitting)