Advanced Genetic Epidemiology
Statistical Workshop

NIDA 2012 AGES WORKSHOP
October 22-26, 2012
Overlook

This workshop is designed to provide an overview of advanced statistical methodology for genetic studies of substance use and abuse phenotypes.

It covers analytical methods for twin and family studies, including measurement and phenotyping, development, family processes and GxE interaction.
<table>
<thead>
<tr>
<th>Paradigm</th>
<th>Samples Studied</th>
<th>Method of Inquiry</th>
<th>Scientific Goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Basic genetic epidemiology</td>
<td>Family, twin, and adoption studies</td>
<td>Statistical</td>
<td>To quantify the degree of familial aggregation and/or heritability</td>
</tr>
<tr>
<td>2. Advanced genetic epidemiology</td>
<td>Family, twin, and adoption studies</td>
<td>Statistical</td>
<td>To explore the nature and mode of action of genetic risk factors</td>
</tr>
<tr>
<td>3. Gene finding</td>
<td>High-density families, trios, case-control samples</td>
<td>Statistical</td>
<td>To determine the genomic location and identity of susceptibility genes</td>
</tr>
<tr>
<td>4. Molecular genetics</td>
<td>Individuals</td>
<td>Biological</td>
<td>To identify critical DNA variants and trace the biological pathways from DNA to disorder</td>
</tr>
<tr>
<td>Heritability</td>
<td>Psychiatric Disorders</td>
<td>Other Important Familial Traits</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------</td>
<td>-------------------------------</td>
<td></td>
</tr>
<tr>
<td>~zero</td>
<td></td>
<td>Language Religion</td>
<td></td>
</tr>
<tr>
<td>20-40%</td>
<td>Anxiety disorders, Depression, Bulimia, Personality Disorders</td>
<td>Myocardial Infarction, Normative Personality, Breast Cancer, Hip Fracture</td>
<td></td>
</tr>
<tr>
<td>40-60%</td>
<td>Alcohol Dependence Drug Dependence</td>
<td>Blood Pressure, Asthma, Plasma cholesterol, Prostate Cancer, Adult-onset diabetes</td>
<td></td>
</tr>
<tr>
<td>60-80%</td>
<td>Schizophrenia Bipolar Illness</td>
<td>Weight, Bone Mineral Density</td>
<td></td>
</tr>
<tr>
<td>80-100%</td>
<td>Autism</td>
<td>Height, Total Brain Volume</td>
<td></td>
</tr>
</tbody>
</table>
Twin study

- Twins are a valuable source for observation because their genotypes and family environments tend to be similar.

- Monozygotic (MZ) or "identical" twins, share nearly 100% of their genetic polymorphisms, which means that most variation in pairs' traits is due to their unique experiences.

- Dizygotic (DZ) or "fraternal" twins share only about 50% of their polymorphisms. Fraternal twins are helpful to study because they tend to share many aspects of their environment by virtue of being born in the same time and place.
Quick review

• Structural Equation Modeling
• Path Diagram
• ACE Model
• R & OpenMX
Structural Equation Modeling

• Structural Equation Modeling is a very general, and powerful multivariate analysis technique allow both confirmatory and exploratory modeling, meaning they are suited to both theory testing and theory development.

• Factor analysis, path analysis and regression all represent special cases of SEM.
Structural Equation Modeling

• Among the strengths of SEM is the ability to construct latent variables: variables which are not measured directly, but are estimated in the model from several measured variables.

http://en.wikipedia.org/wiki/Structural_equation_modeling
The Basic Idea Behind Structural Modeling

• One of the fundamental ideas taught in intermediate applied statistics courses is the effect of additive and multiplicative transformations on a list of numbers. Students are taught that, if you multiply every number in a list by some constant K, you multiply the mean of the numbers by K. Similarly, you multiply the standard deviation by the absolute value of K.
Structural Equation Modeling

• For example, suppose you have the list of numbers 1,2,3. These numbers have a mean of 2 and a standard deviation of 1. Now, suppose you were to take these 3 numbers and multiply them by 4. Then the mean would become 8, and the standard deviation would become 4, the variance thus 16.
Structural Equation Modeling

• The point is, if you have a set of numbers X related to another set of numbers Y by the equation $Y = 4X$, then the variance of Y must be 16 times that of X, so you can test the hypothesis that Y and X are related by the equation $Y = 4X$ indirectly by comparing the variances of the Y and X variables.
Structural Equation Modeling

• This idea generalizes, in various ways, to several variables inter-related by a group of linear equations. The rules become more complex, the calculations more difficult, but the basic message remains the same -- you can test whether variables are interrelated through a set of linear relationships by examining the variances and covariances of the variables.

http://www.statsoft.com/textbook/structural-equation-modeling/
Path Diagram

• Path Diagrams play a fundamental role in structural modeling. Path diagrams are like flowcharts.

• They show variables interconnected with lines that are used to indicate causal flow.

• Most structural equation models can be expressed as path diagrams.
Path Diagram

• Consider the classic linear regression equation

\[ y = ax + e \]

• Any such equation may be represented in a path diagram as follows:

http://www.statsoft.com/textbook/structural-equation-modeling/
ACE model

• Typically these three components are called:
  
  A  (additive genetics)
  C  (common environment)
  E  (unique environment)

It is also possible to examine non-additive genetics effects (often denoted D for dominance)
ACE model

• Monozygotic (identical - MZ) twins raised in a family share both 100% of their genes, and all of the shared environment. Any differences arising between them in these circumstances are random (unique).

• The correlation we observe between identical twins provides an estimate of $A + C$. 
ACE model

- Dizygous (DZ) twins have a common shared environment, and share on average 50% of their genes:
- so the correlation between fraternal twins is a direct estimate of $\frac{1}{2}A + C$. 
ACE model

• If $r$ is the correlation observed for a particular trait, then:

  • $r_{mz} = A + C$
  • $r_{dz} = \frac{1}{2}A + C$

• Where $r_{mz}$ and $r_{dz}$ are simply the correlations of the trait in identical and fraternal twins respectively.

http://en.wikipedia.org/wiki/Twin_study
R & OpenMX

• What is OpenMx?

• OpenMx is free and open source software for use with R that allows estimation of a wide variety of advanced multivariate statistical models.

• OpenMx consists of a library of functions and optimizers that allow you to quickly and flexibly define Structural equation modeling (SEM) model and estimate parameters given observed data.

http://openmx.psyc.virginia.edu/
R & OpenMX

• **OpenMx** can be used by those who think in terms of *path models* or by those who prefer to specify models in terms of matrix algebra.
Path Model Specification

• Here is a path diagram for a one factor path model with five indicators. Beside it is an R script using OpenMx path modeling commands to read the data from disk, create the one factor model, fit the model to the observed covariances, and print a summary of the results.
require(OpenMx)
data(demoOneFactor)
manifests <- names(demoOneFactor)
latts <- c("G")
factorModel <- mxModel("One Factor",
  type="RAM",
  manifestVars = manifests,
  latentVars = latts,
  mxPath(from=latts, to=manifests),
  mxPath(from=manifests, arrows=2),
  mxPath(from=latts, arrows=2,
    free=FALSE, values=1.0),
  mxData(cov(demoOneFactor),
    type="cov",
    numObs=500))
summary(mxRun(factorModel))
Matrix Model Specification

• **OpenMx** can also specify models in terms of matrix algebra. On the left is an equation for the same one factor path model with five indicators. Beside it is an R script using **OpenMx** matrix modeling commands to read the data from disk, create the one factor model, fit the model to the observed covariances, and print a summary of the results.
Matrix Model Specification

R = ALA' + U

data(demoOneFactor)
factorModel <- mxModel("One Factor",
  mxMatrix("Full", 5, 1, values=0.2,
    free=TRUE, name="A"),
  mxMatrix("Symm", 1, 1, values=1,
    free=FALSE, name="L"),
  mxMatrix("Diag", 5, 5, values=1,
    free=TRUE, name="U"),
  mxAlgebra(A %*% L %*% t(A) + U,
    name="R"),
  mxMLObjective("R", dimnames =
    names(demoOneFactor)),
  mxData(cov(demoOneFactor),
    type="cov", numObs=500))
summary(mxRun(factorModel))
Multivariate Genetic Analysis - Question

**Univariate Analysis:** What are the contributions of additive genetic, dominance/shared environmental and unique environmental factors to the variance?

**Bi/Multivariate Analysis:** What are the contributions of genetic and environmental factors to the covariance between two traits? What makes sets of variables correlate or co-vary, comorbid?
Univariate ACE model

http://www.slideshare.net/devenvaija09/bivariate
Expected Covariance Matrices

\[ \Sigma_{MZ} = \begin{bmatrix} a^2+c^2+e^2 & a^2+c^2 \\ a^2+c^2 & a^2+c^2+e^2 \end{bmatrix} \quad 2 \times 2 \]

\[ \Sigma_{DZ} = \begin{bmatrix} a^2+c^2+e^2 & .5a^2+c^2 \\ .5a^2+c^2 & a^2+c^2+e^2 \end{bmatrix} \quad 2 \times 2 \]
Bivariate Questions I

- Univariante Analysis: What are the contributions of additive genetic, dominance/shared environmental and unique environmental factors to the variance?
- Bivariate Analysis: What are the contributions of genetic and environmental factors to the covariance between two traits?
Two Traits
Bivariate Questions II

- Two or more traits can be correlated because they share common genes or common environmental influences
  - e.g. Are the same genetic/environmental factors influencing the traits?
- With twin data on multiple traits it is possible to partition the covariation into its genetic and environmental components
- Goal: to understand what factors make sets of variables correlate or co-vary
### Bivariate Twin Data

<table>
<thead>
<tr>
<th></th>
<th>individual twin</th>
</tr>
</thead>
<tbody>
<tr>
<td>within</td>
<td>within</td>
</tr>
<tr>
<td></td>
<td>variance</td>
</tr>
<tr>
<td>between</td>
<td>twin covariance</td>
</tr>
<tr>
<td></td>
<td>trait covariance</td>
</tr>
<tr>
<td></td>
<td>cross-trait</td>
</tr>
<tr>
<td></td>
<td>twin covariance</td>
</tr>
</tbody>
</table>
## Bivariate Twin Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$Y_1$</th>
<th>$X_2$</th>
<th>$Y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$V_{X_1}$</td>
<td>$C_{X_1Y_1}$</td>
<td>$C_{X_1X_2}$</td>
<td>$C_{X_1Y_2}$</td>
</tr>
<tr>
<td>$Y_1$</td>
<td>$C_{Y_1X_1}$</td>
<td>$V_{Y_1}$</td>
<td>$C_{Y_1X_2}$</td>
<td>$C_{Y_1Y_2}$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$C_{X_2X_1}$</td>
<td>$C_{X_2Y_1}$</td>
<td>$V_{X_2}$</td>
<td>$C_{X_2Y_2}$</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>$C_{Y_2X_1}$</td>
<td>$C_{Y_2Y_1}$</td>
<td>$C_{Y_2X_2}$</td>
<td>$V_{Y_2}$</td>
</tr>
</tbody>
</table>
Genetic Correlation

\[ A_1 \xrightarrow{r_g} A_2 \xrightarrow{1} X_1 \xrightarrow{1} Y_1 \]

\[ A_1 \xleftarrow{r_g} A_2 \xleftarrow{1} X_2 \xleftarrow{1} Y_2 \]
Alternative Representations
Cholesky Decomposition
More Variables
Bivariate AE Model
### MZ Twin Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$Y_1$</th>
<th>$X_2$</th>
<th>$Y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$a_{11}^2 + e_{11}^2$</td>
<td>$a_{11}^2$</td>
<td>$a_{11}^2$</td>
<td>$a_{11}^2$</td>
</tr>
<tr>
<td>$Y_1$</td>
<td>$a_{21}a_{11} + e_{21}e_{11}$</td>
<td>$a_{22}^2 + a_{21}^2 + e_{22}^2 + e_{21}^2$</td>
<td>$a_{21}a_{11}$</td>
<td>$a_{22}^2 + a_{21}^2$</td>
</tr>
<tr>
<td>$X_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### DZ Twin Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$Y_1$</th>
<th>$X_2$</th>
<th>$Y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$a_{11}^2 + e_{11}^2$</td>
<td>.5$a_{11}^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_1$</td>
<td>$a_{21}^2 * a_{11} + e_{21}^2 * e_{11}$</td>
<td>$a_{22}^2 + a_{21}^2 + e_{22}^2 + e_{21}^2$</td>
<td>.5$a_{21}^2 * a_{11}$</td>
<td>.5$a_{22}^2 + .5a_{21}^2$</td>
</tr>
<tr>
<td>$X_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Within-Twin Covariances [Mx]

\[ A = X^*X' \]

\[ \Sigma A = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix} \times \begin{bmatrix} a_{11} & a_{21} \\ 0 & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11}^2 & a_{11}a_{21} \\ a_{21}a_{11} & a_{22}^2 + a_{21}^2 \end{bmatrix} \]
Within-Twin Covariances

\[ \Sigma A = \begin{bmatrix} a_{11}^2 & a_{11}a_{21} \\ a_{21}a_{11} & a_{22}^2 + a_{21}^2 \end{bmatrix} \]

\[ \Sigma E = \begin{bmatrix} e_{11}^2 & e_{11}e_{21} \\ e_{21}e_{11} & e_{22}^2 + e_{21}^2 \end{bmatrix} \]

\[ \Sigma P = \Sigma A + \Sigma E = \begin{bmatrix} a_{11}^2 + e_{11}^2 & a_{11}a_{21} + e_{11}e_{21} \\ a_{21}a_{11} + e_{21}e_{11} & a_{22}^2 + a_{21}^2 + e_{22}^2 + e_{21}^2 \end{bmatrix} \]
Cross-Twin Covariances

\[
\begin{align*}
\text{MZ} & \quad \Sigma A = \\
& = \begin{bmatrix}
 a_{11}^2 & a_{11}a_{21} \\
 a_{21}a_{11} & a_{22}^2 + a_{21}^2 \\
\end{bmatrix} \\
\text{DZ} & \quad .5@ \Sigma A = \\
& = \begin{bmatrix}
 .5a_{11}^2 & .5a_{11}a_{21} \\
 .5a_{21}a_{11} & .5a_{22}^2 + .5a_{21}^2 \\
\end{bmatrix}
\end{align*}
\]
Cross-Trait Covariances

- Within-twin cross-trait covariances imply common etiological influences
- Cross-twin cross-trait covariances imply familial common etiological influences
- MZ/DZ ratio of cross-twin cross-trait covariances reflects whether common etiological influences are genetic or environmental
### Univariate Expected Covariances

\[
\Sigma_{MZ} = \begin{bmatrix}
  a^2 + c^2 + e^2 & a^2 + c^2 \\
  a^2 + c^2 & a^2 + c^2 + e^2
\end{bmatrix} \quad 2 \times 2
\]

\[
\Sigma_{DZ} = \begin{bmatrix}
  a^2 + c^2 + e^2 & \frac{1}{2}a^2 + c^2 \\
  \frac{1}{2}a^2 + c^2 & a^2 + c^2 + e^2
\end{bmatrix} \quad 2 \times 2
\]
Univariate Expected Covariances II

$\Sigma \text{ MZ} =$

$$
\begin{array}{c|c}
\Sigma A + \Sigma C + \Sigma C & \Sigma A + \Sigma C \\
\hline
\Sigma A + \Sigma C & \Sigma A + \Sigma C + \Sigma C
\end{array}
$$

$2 \times 2$

$\Sigma \text{ DZ} =$

$$
\begin{array}{c|c}
\Sigma A + \Sigma C + \Sigma C & .5@\Sigma A + \Sigma C \\
\hline
.5@\Sigma A + \Sigma C & \Sigma A + \Sigma C + \Sigma C
\end{array}
$$

$2 \times 2$
Bivariate Expected Covariances

$\Sigma MZ =$

\[
\begin{array}{cc}
\Sigma A + \Sigma C + \Sigma C & \Sigma A + \Sigma C \\
\Sigma A + \Sigma C & \Sigma A + \Sigma C + \Sigma C \\
\end{array}
\]

$4 \times 4$

$\Sigma DZ =$

\[
\begin{array}{cc}
\Sigma A + \Sigma C + \Sigma C & .5@\Sigma A + \Sigma C \\
.5@\Sigma A + \Sigma C & \Sigma A + \Sigma C + \Sigma C \\
\end{array}
\]

$4 \times 4$
Practical Example I

- Dataset: MCV-CVT Study
- 1983-1993
- BMI, skinfolds (bic, tri, calf, sil, ssc)
- Longitudinal: 11 years
- N MZFY: 107, DZF: 60
Practical Example II

- Dataset: NL MRI Study
- 1990’s
- Working Memory, Gray & White Matter

- $N_{MZFY}$: 68, $DZF$: 21
Cholesky decomposition

• A typical starting point in bivariate and multivariate analysis is the **Cholesky decomposition**.
Cholesky decomposition

• Given a symmetric positive definite matrix $A$, the Cholesky decomposition is an upper triangular matrix $U$ with strictly positive diagonal entries such that

$$A = U^T U.$$
Cholesky decomposition

• The most commonly used multivariate technique in the Classical Twin Design is **Cholesky decomposition**.

• The Cholesky is a method of triangular decomposition where the **first variable** \((y_1)\) is assumed to be **caused by a latent factor** \((\eta_1)\) that can explain the variance in remaining variables \((y_2, \ldots, y_n)\) and so on.
Cholesky decomposition

Figure 1  Multivariate Cholesky triangular decomposition, $y_1, \ldots, y_n$ = observed phenotypic variables, $\eta_1-\eta_n$ = latent factors
Cholesky decomposition

• The expected variance-covariance matrix in the Cholesky decomposition is parameterized in terms of \textit{n latent factors} (Where \(n\) is the number of variables).

• All variables load on the first latent factor, \(n-1\) variables load on the second factor and so on, the \textit{final variable} loads on the \textit{\(n^\text{th}\) latent factor} only. Each source of phenotypic variation (i.e A, C or D, E) is parameterized in the same way.
Cholesky decomposition

• Therefore, the full factor Cholesky does not distinguish between common factor and specific factor variance and does not estimate a specific factor effect for any variable except the last.