

MEETING THE CHALLENGE OF CLUSTERED AND MULTILEVEL DATA

Keith E. Muller

Department of Health Outcomes and Policy
University of Florida, KMuller@ufl.edu
www.health-outcomes-policy.ufl.edu/muller

Supported in part by NIAAA R01-AA013458-01,
NHLBI R01-HL091005,
NIDCR U54-DE019261, and
NIDCR R01-DE020832-01A1

Disclosures: Muller has been a paid consultant to SAS;
his wife worked for SAS for 5 years.

- 1. Recognizing the Challenge**
 - 2. Achieving Robust Covariance in Mixed Models**
 - 3. Covariance Patterns for Multilevel Data**
- Bibliography

1. Recognizing the Challenge

1.1 Motivation

Cluster and multilevel sampling can help:

More sensitivity (power)

Lower costs (recruit fewer participants)

Cluster and multilevel sampling can hinder:

Improper analysis usually underestimates variance and inflates type I error rate.

*Need to know when and how to use
to gain the advantages and avoid the problems.*

1.2 Definitions

All of Statistics: estimation and
inference (testing)

Intervention study:

modeling relationships (estimation) and
testing hypotheses

Response = f (predictor)

With random assignment to treatment,

Dependent = f (independent)

Statistical model:

Response = f (predictor) + error

Statistical model:

$$\text{response} = f(\text{predictor}) + \text{error}$$

Responses: Y variables

Predictors: X variables

Errors: E

$$Y = f(X) + E$$

Error Variance:

“job security for statisticians”

1.3 Diagnostic Questions

Question 1: How Many Variables?

# responses	# predictors	Model
1	1	univariate
1	many	multi-variable
many	1	multivariate
many	many	multivariate

Same Y , many times: Multivariate,
and also Repeated Measures (REPM)

Measures between sampling units
(person, machine, hospital . . .) are independent.

Measures within (repeated measures:
"Time," limb, organ, . . .) are not independent
(correlated).

Many distinct Y 's, many times:
doubly multivariate, collections of repeated measures

Independent observations

versus {
Multivariate (many Y 's)
Repeated Measures
Correlated observations
Non-independent observations }

Emphasize: statistics different!

Improper analysis can severely bias results

Theory same for multivariate and REPM while interpretation and analysis differ.

Question 2: Variable Types?

Scale Data (Error Distribution)

Nominal: Dichotomous
 Polychotomous
 Ordered Categorical

Ordinal Order with infinitely many values.

Interval } Continuous { Gaussian
Ratio } { Non-Gaussian

Maximize power by avoiding categorical data or
categorizing continuous data.

Inaccuracy in small sample inference for most methods.

Question 3: Sampling Pattern?

Design	#Times	# Indep. Sampling Units	Timing
Cross Sectional	one	N	--

Everything else: not cross sectional,
multivariate in some sense

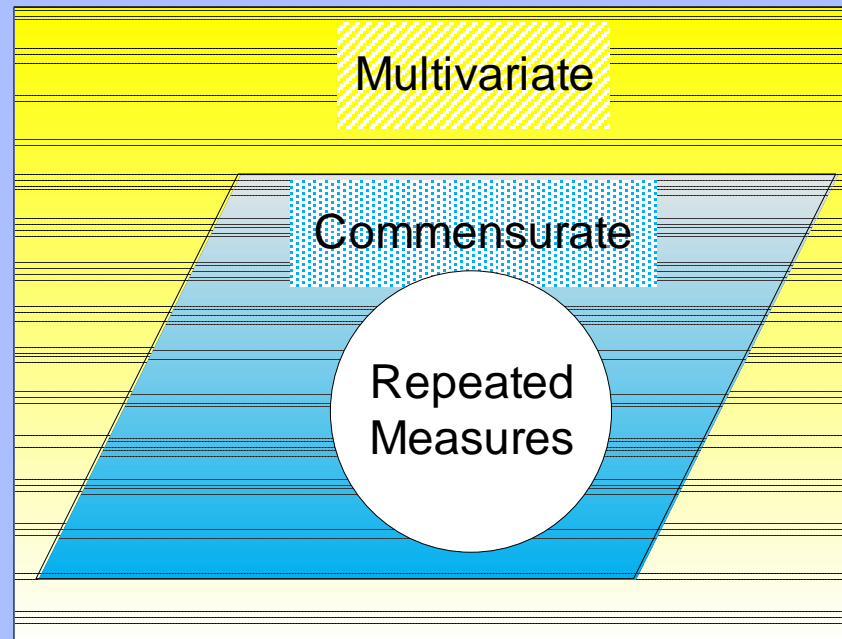


Figure 1. Categories of Multivariate Data

All variables measured in same units, *commensurate*.

What is the independent sampling unit,
in contrast to the observational unit?

observations = # ISU * # Times

N = # independent sampling units (ISU)

p_i = # observations for independent sampling unit i

n = total # observations = $\sum_{i=1}^N p_i$

In a cross sectional design $n = N$ because $p_i \equiv 1$

Table 1. Repeated Measures Sampling Patterns

Type	# Obs <i>per ISU</i>	#ISU	Timing
REPM	> 1	N	consistent
Crossover	p	N	alternating
Longitudinal	p_i , varies	N	inconsistent
Time series	n	1	regular
Cluster	p_i	# Clust	exchangeable
Survey	p_i	# Clust	exchangeable

Split Plot terminology originated in agriculture

Hierarchical, or multilevel data have "nested" sampling, with combinations of dimensions.

Examples of two or more dimensions of sampling:

teeth within a person

teeth within a person within a clinic

teeth within a person within a clinic

measured repeatedly over *time*

*Question 4: # Independent Sampling Units
per Response Variable?*

1's?

10's?

100's?

1000's?

more?

High Dimension, Low Sample Size (HDLSS):

Examples of more observations than people:

genomics,

transcripomics,

medical imaging

Question 1: How Many Variables?

Question 2: Variable Types?

Question 3: Sampling Scheme?

Question 4: # ISUs per Response?

Answers to questions

allow recognizing the challenge.

Caution:

Usage of some definitions varies widely!

1.4 Analysis Methods

Choosing a Technique

Accurate estimation (means, proportions)?

Defensible inference (type I error rate)?

Property 1: How Many Variables?

Property 2: Variable Types?

Property 3: Sampling Scheme?

Property 4: # Persons per Response?

$5 \times 6 \times 8 \times 5 = 1200$ answers to 4 questions.
Each a potentially distinct analysis.
Answer requires a career, not a lecture!

Every analysis an approximation.
Seek accurate estimation and
defensible inference (tests).

Answers changing every few years
due to advances in computing.

“Being a statistician means
never having to say you're certain.”
(ASA T-shirt)

Shop smart, learn limitations.

Report any limitations in publications.

A bothersome scenario:

statisticians propose new technique,
prove good for large samples,
run simulations for easy cases for estimation,
give large N (100's, 1000's) study example,
use large sample inference (Z test, χ^2 test),
declare ready for general use.

Small (1's, 10's) sample tests not defensible;
possibly acceptable, but unknown.

Extreme versions often seen in HDLSS.

Seems to invoke theatre response:
a willing suspension of disbelief.

Accurate small sample tests not available for many interesting REPM models, even assuming Gaussian errors.

Worst problem for non-Gaussian data.

Availability of software to fit a model, even in major packages, does not guarantee method defensible for small N studies.

1. Recognizing The Challenge

- ✓ 1.1 Motivation: *Special Handling*
- ✓ 1.2 Definitions
- ✓ 1.3 Diagnosis: *Answer 4 Questions*
- ✓ 1.4 Analysis Methods

2. Achieving Robust Covariance In Mixed Models

- 2.1 Inference Robust to Under-Specified Covariance?
- 2.2 Large N Robust to Under-Specified Covariance?
- 2.3 Strategies for Robust Inference in Fixed Effects

3. Covariance Patterns for Multilevel Data

2. Achieving Robust Covariance in Mixed Models

Gurka, Muller and Edwards (2011)

Avoiding bias in mixed model inference for fixed effects, *Statistics in Medicine*, in press.

1. Mixed model myth: fixed effects inference robust to under-specified covariance model (**BUSTED**)
2. Mixed model myth refined:
large sample makes inference robust to under-specified covariance model (**BUSTED**)
3. Strategies that help achieve robust inference:
Condition the data
Choose sufficiently complex and estimable covariance
Choose test statistic and approximation carefully

2.1 Is Fixed Effect Inference Robust to Under-Specified Covariance?

Study Conditions: Best Case Scenario

Complete, balanced data (no missing values),
no covariates changing with time (except time)
Gaussian data with homogeneity between groups

A multivariate model dressed up as a mixed model.

(chapter 12, Muller and Stewart, 2006, *Linear Model Theory; Univariate, Multivariate, and Mixed Models*)

Σ_i = population covariance of $\{y_{ij}\}$ response values
for independent sampling unit (person) $i \in \{1, \dots, N\}$

Kenward-Roger approximation for Wald test *falsely*
assuming a single random effect (Σ_i model).

KR clear winner in all published simulations.

Proof of Claim

1. Special case allows reducing $\hat{\beta}$, $\hat{\Sigma}_i$ and F_W statistic to closed forms.
2. Assume Σ_i unstructured while data analysis *falsely* assumes a single random effect, which implies $\hat{\Sigma}_i$ being compound symmetric.
3. Derive F_W as a ratio of positive quadratic forms, convert to general quadratic form with \pm coefficients.
4. Davies' algorithm computed exact test size to give counterexamples in small and large finite samples.
5. Use simulations to check enumerations.

Table 2. Exact Type I Error ($\times 100$)
 Fixed Effect Interaction, $\alpha = 0.05$,
 Assuming Compound Symmetry, Complete Data

Study Size	True Covariance, $\Sigma_i = \mathcal{V}(\mathbf{y}_i)$			
	1,IID	1,AR	2,IID	2,AR
$p = 5, N = 50$	5.0	9.0	9.8	13.7
$p = 5, N = 100$	5.0	9.0	9.8	13.7
$p = 5, N = \infty$				
$p = 10, N = 50$	5.0	10.9	14.5	19.7
$p = 10, N = 100$	5.0	11.0	14.5	19.7
$p = 10, N = \infty$				

1,IID: Randm intercpt, i.i.d. within error; cmpnd symtry

1,AR: Randm intercept, AR(1) within error

2,IID: Randm intercpt & slope, i.i.d. within error

2,AR: Randm intercpt & slope, AR(1) within error

2.2 Large N Robust to Under-Specified Covariance?

Derivation for $N \rightarrow \infty$, Fixed $\{p_i\}$

1. Compute term-by-term limits of quadratic form components for $N = \#$ ISU (Independent Sampling Unit, person) $\rightarrow \infty$ and fixed $p_i = p$ (# observations per ISU, # repeated measures)
2. In the ratio form, $F_W = Q_1/Q_2$, the denominator Q_2 converges to a constant, while the numerator Q_1 does not depend on N , so goes nowhere.

Same inflation of type I error results as with finite N .

Table 3. Exact Type I Error ($\times 100$)
 Fixed Effect Interaction, $\alpha = 0.05$,
 Assuming Compound Symmetry, Complete Data

Study Size	True Covariance, $\Sigma_i = \mathcal{V}(\mathbf{y}_i)$			
	1,IID	1,AR	2,IID	2,AR
$p = 5, N = 50$				
$p = 5, N = 100$				
$p = 5, N = \infty$	5.0	9.0	9.8	13.7
$p = 10, N = 50$				
$p = 10, N = 100$				
$p = 10, N = \infty$	5.0	11.0	14.5	19.7

1,IID: Randm intercpt, i.i.d. within error; cmpnd symtry

1,AR: Randm intercpt, AR(1) within error

2,IID: Randm intercpt & slope, i.i.d. within error

2,AR: Randm intercpt & slope, AR(1) within error

Why Does $N = \infty$ Give No Help?

In the simplest case of a scalar parameter

$$\begin{aligned} F_W &= Q_1/Q_2 \\ &= \hat{\theta}^2 / \hat{\mathcal{V}}(\hat{\theta}) \end{aligned}$$

However $E[\hat{\mathcal{V}}(\hat{\theta})] \neq \mathcal{V}(\hat{\theta})$ at finite N or as $N \rightarrow \infty$

Biased and inconsistent; Q_2 converges to the *wrong* constant.

*Simulation Results for Missing Data:
Will Uglier Data Make the Results Prettier?*

Figure 2. Observed Type I Error $\times 100$, Fixed Effect Interaction, $\alpha = 0.05$, 20% Missing Completely at Random, 5 obs per Person, 10,000 Replications

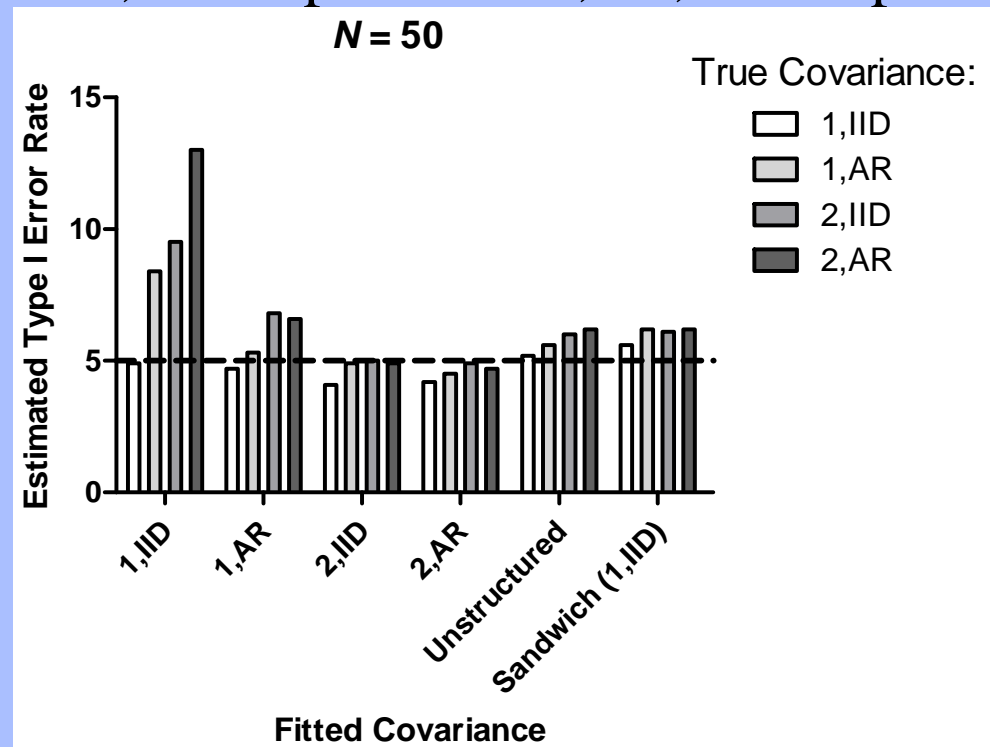
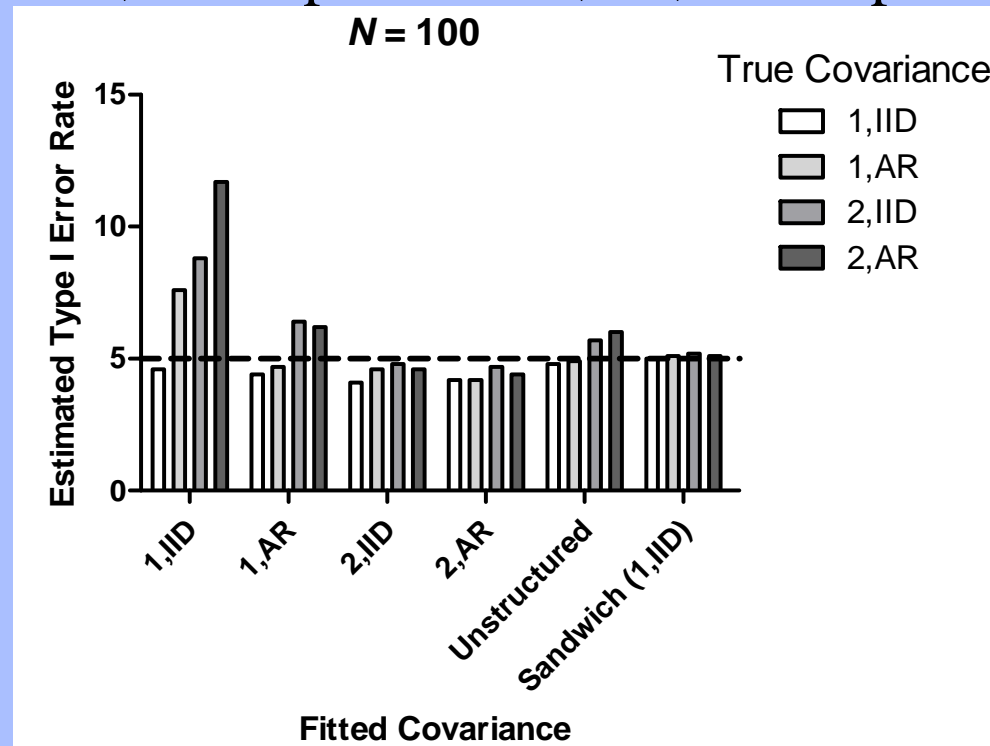


Figure 3. Observed Type I Error $\times 100$, Fixed Effect Interaction, $\alpha = 0.05$, 20% Missing Completely at Random, 5 obs per Person, 10,000 Replications



Review

2.1 Mixed model myth:

fixed effects inference robust to under-specified covariance model

✓**[BUSTED]**

2.2 Mixed model myth refined:

large sample makes inference robust to under-specified covariance model

✓**[BUSTED]**

2.3 Strategies to achieve robust inference for fixed effects

2.3 Strategies for Robust Inference in Fixed Effects

Strategy 1. Use Five-Step Procedure for Model Fitting

Cheng, Edwards, Maldonado-Molina, Komro, and Muller (2010) Real longitudinal data analysis for real people: building a good enough mixed model, *Statistics in Medicine*, **29**, 504-520.

The five-step procedure for model fitting generalizes approach for univariate models in Chapter 11 of

Muller and Fetterman (2002) *Regression and ANOVA: An Integrated Approach Using SAS[®] Software*. Cary, NC: SAS Institute.

Strategy 2. Improve the Condition of the Data Center,
scale, and
full-rank code all predictors (\mathbf{X}, \mathbf{Z})
to improve the chances of convergence,
computing speed, and
numerical accuracy.

Strategy 3. Further Condition the Data

Round the time values to the roughest level scientifically tolerable.

Simplify the model by "augen-analysis:" use your eyes and think.

Example: for growth or death process, use linear+quadratic+cubic and drop higher order terms.

Change goals to align the question asked of the data with the limits of the data.

Extreme version: declare the data inadequate to answer any questions of interest.

Strategy 4. Eliminate Computational Problems

Apply computational and assumption diagnostics from univariate linear models to detect and solve computational problems and improve the chances of convergence, computing speed, and numerical accuracy.

Strategy 5. Choose Test and Approximation Carefully

Classic questions: what controls test size, even in small samples?

if controls test size, what power available?

what robustness to assumptions?

Kenward-Roger if using Gaussian mixed model.

Seek two-moment approximations (Bartlett correction for χ^2 ,

F approximation in some cases).

Deviance statistic perhaps?

Review

1. Mixed model myth:
fixed effects inference robust to under-specified
covariance model
✓[**BUSTED**]
2. Mixed model myth refined:
large sample makes inference robust to under-
specified covariance model
✓[**BUSTED**]
3. Practical strategies can help achieve robust inference
for fixed effects in mixed models.

Conclusions

Smart model fitting requires choosing a sufficiently complex and estimable covariance structure.

The strategies help fit an adequately complex covariance model,
a crucial need for defensible inference.

Success takes time, work, and commitment to good statistical practice.

1. Recognizing The Challenge

- ✓ 1.1
- ✓ 1.2
- ✓ 1.3
- ✓ 1.4

2. Achieving Robust Covariance In Mixed Models

- ✓ 2.1 Inference Robust to Under-Specified Covariance?
- ✓ 2.2 Large N Robust to Under-Specified Covariance?
- ✓ 2.3 Strategies for Robust Inference in Fixed Effects

3. Covariance Patterns for Multilevel Data

- 3.1 Separate Model for Each Level
- 3.2 LEAR a New Tool for Longitudinal Data
- 3.3 Direct Product Models Solve Many Problems

3. Covariance Patterns for Multilevel Data

3.1 Separate Model for Each Level

Each level demands a specific covariance pattern, with specific parameter values.

Person within clinic? Exchangeable, so compound symmetry, i. e. equal variance and equal correlation.

Teeth? Not exchangeable, unstructured covariance?

Covariance across time? Unstructured?

3.2 LEAR, a New Tool for Longitudinal Data

Linear Exponential Autoregressive (LEAR),
a generalization of AR

Simpson, Edwards, Muller, Sen, and Styner (2010)

$$\rho_{ei;jk} = \begin{cases} 1 + [(d(t_{ij}, t_{ik}) - 1)\delta_e / (D - 1)] & j \neq k \\ \rho_e & \\ 1 & j = k \end{cases}$$

$d(t_{ij}, t_{ik})$ = distance between times or locations,

D a computational flexibility *constant*, by default the maximum number of distance units,

ρ_e is the correlation between observations separated by one unit of time or distance

δ_e is the decay speed, $0 \leq \rho_e < 1$, $0 \leq \delta_e$, and $D > 1$.

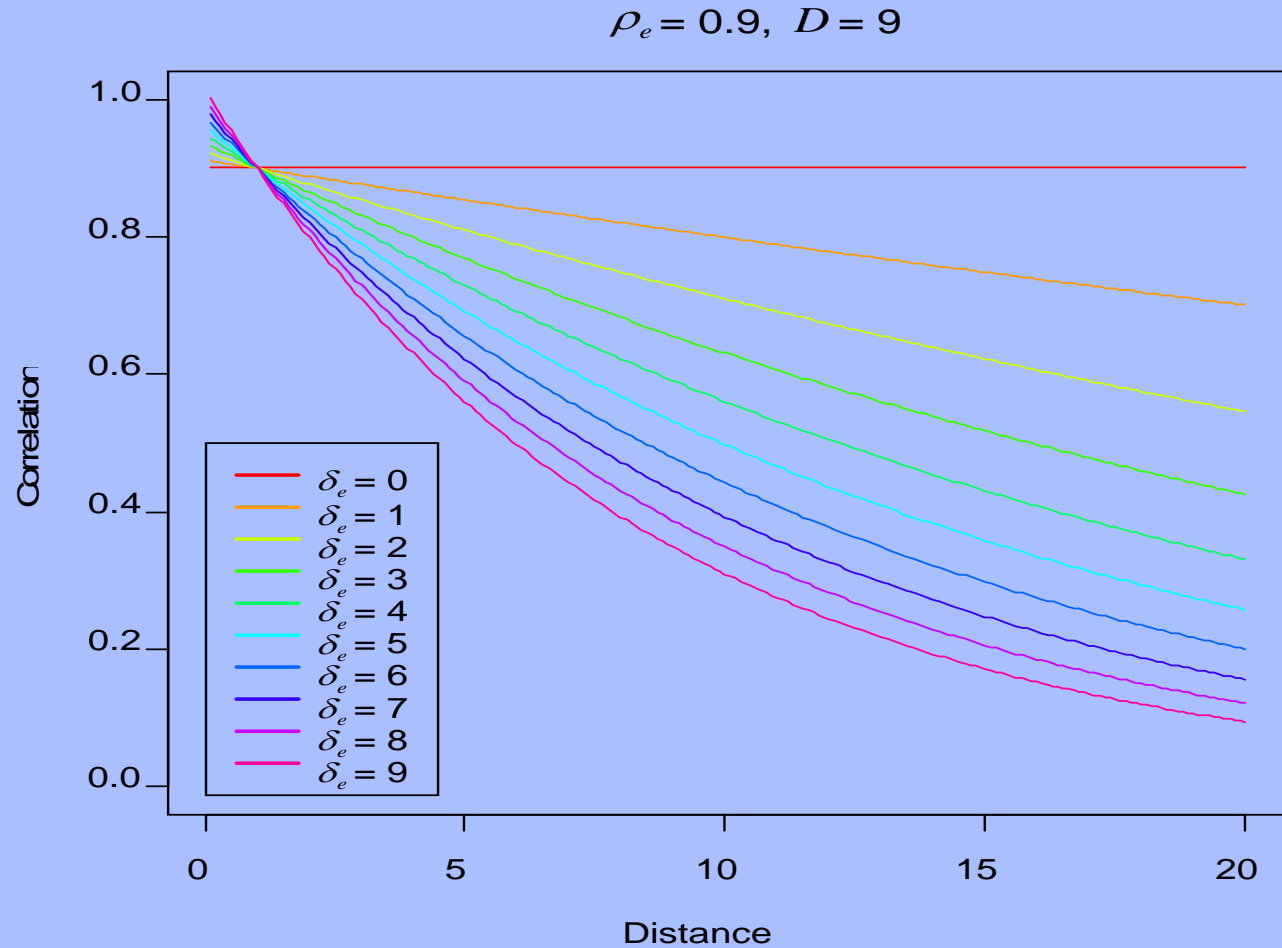


Figure 4. Correlation patterns varying δ_e , constant ρ_e and D ; $\delta_e = 8$ for AR(1), $\delta_e = 0$ for compound symmtry.

$$\delta_e = 1, D = 9$$

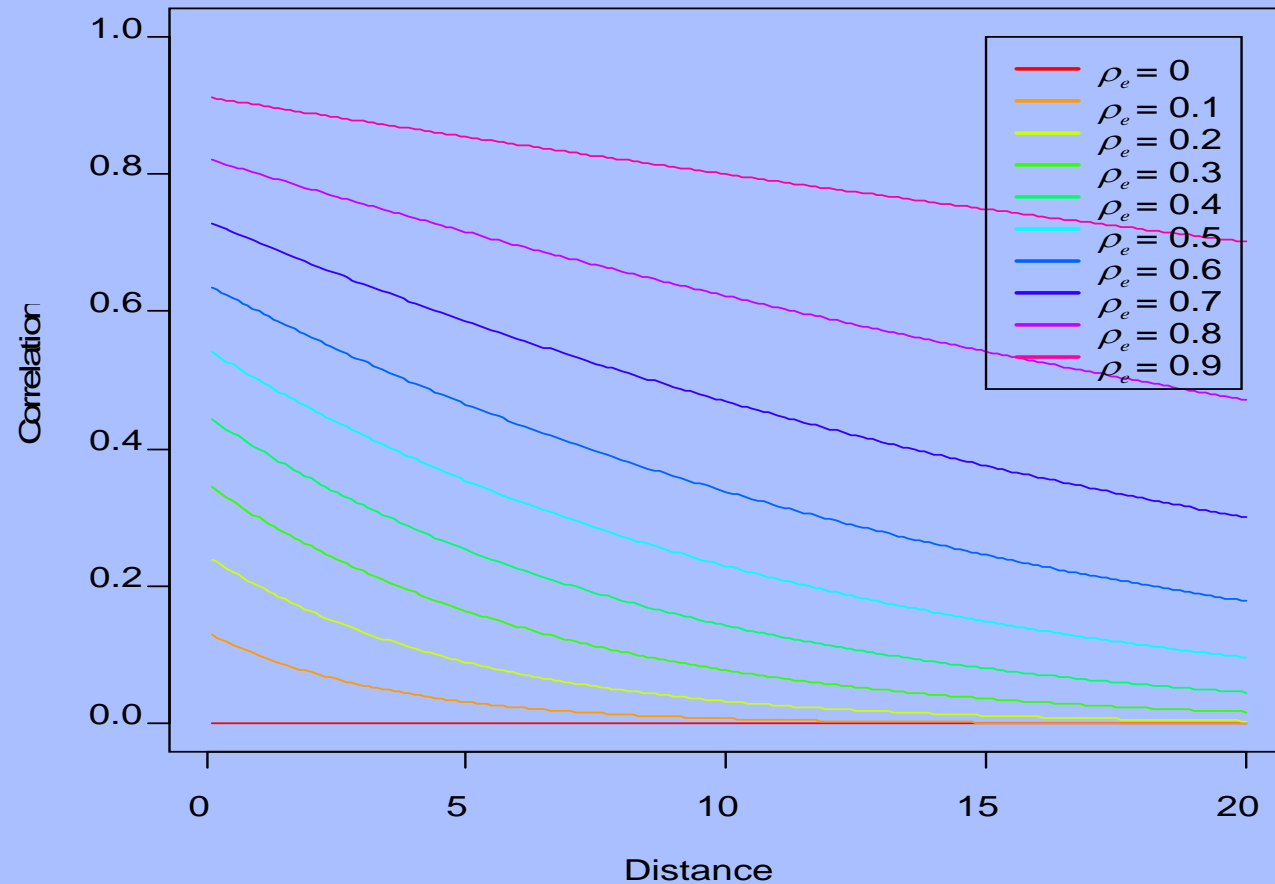


Figure 5. Correlation patterns obtained by varying ρ_e keeping δ_e and D constant.

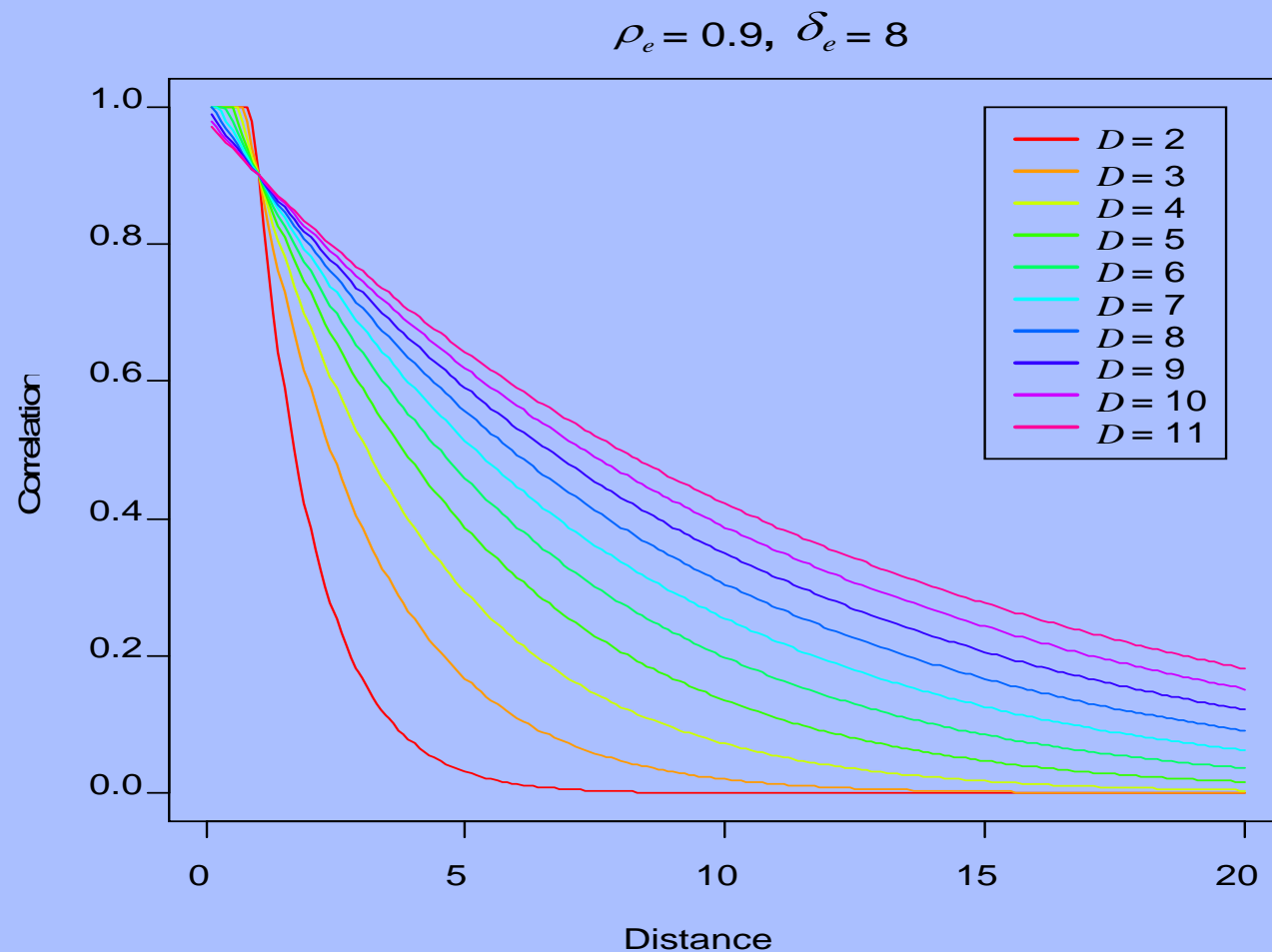


Figure 6. Correlation patterns obtained by varying D keeping δ_e and ρ_e constant. $D = 9$ give an AR(1).

3.3 Direct Product Models Solve Many Problems

Direct Product matrix multiplication, also known as Kronecker product:

$$\begin{aligned} \mathbf{A} \otimes \mathbf{B} &= \{a_{ij}\mathbf{B}\} \\ &= \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1c}\mathbf{B} \\ a_{21}\mathbf{B} & \ddots & \vdots \\ \vdots & & \vdots \\ a_{r1}\mathbf{B} & \cdots & a_{rc}\mathbf{B} \end{bmatrix} \end{aligned}$$

Teeth (unstructured covariance)
within person (compound symmetry),

$$\Sigma_{\text{CS}} \otimes \Sigma_{\text{UN}} = \begin{bmatrix} 1 \cdot \Sigma_{\text{UN}} & \cdots & \rho \cdot \Sigma_{\text{UN}} \\ \rho \cdot \Sigma_{\text{UN}} & \ddots & \vdots \\ \vdots & & \vdots \\ \rho \cdot \Sigma_{\text{UN}} & \cdots & 1 \cdot \Sigma_{\text{UN}} \end{bmatrix}$$

Why Should You Care for Such a Fancy Model?

The model is scientifically credible *and*
the model will converge because have only

$d = 2 + p(p + 1)/2$ parameters

$p = 4$ locations (times) implies $d = 12$

10 people per clinic implies 820 parameters for
unstructured covariance model

1. Recognizing The Challenge

- ✓ 1.1 Motivation: Special Handling
- ✓ 1.2 Definitions
- ✓ 1.3 Diagnosis: Answer 4 Questions
- ✓ 1.4 Analysis Methods

2. Achieving Robust Covariance In Mixed Models

- ✓ 2.1 Inference Robust to Under-Specified Covariance?
- ✓ 2.2 Large N Robust to Under-Specified Covariance?
- ✓ 2.3 Strategies for Robust Inference in Fixed Effects

3. Covariance Patterns for Multilevel Data

- ✓ 3.1 Separate Model for Each Level
- ✓ 3.2 LEAR a New Tool for Longitudinal
- ✓ 3.3 Direct Product Models Solve Many Problems