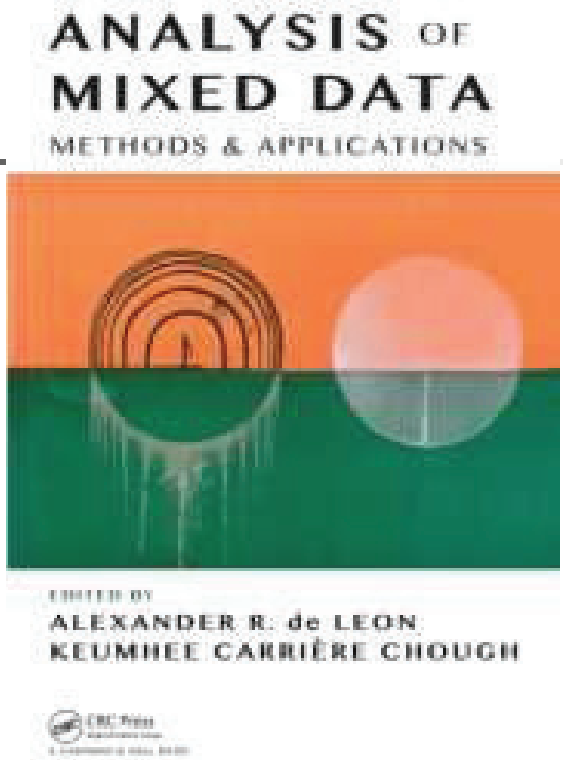


Joint Analysis of Mixed Outcome Data:

Issues and Challenges

By
Keumhee Carriere Chough
University of Alberta
Edmonton, Alberta, Canada

OUTLINE

- Introduction
- Motivations
- History of development for mixed data methods
  - Direct approaches
    - CGD (GLOM, CGCM), GMDM, Copula
  - Indirect approaches
- Further research
- Remarks

Introduction

- Data collection
  - Often come with complex dependence structures and types
  - Require non-standard aproaches to analysis
  - Computationally intensive
- Mixed Data
  - Continuous
  - Ordinal
  - Categorical

Motivating Examples

- Developmental toxicity study of Ethylene glycol in mice (1985).

Table 8.1 *Summary statistics of EG mice data.*

Dose (g/kg)	Number of dams	Number of live fetuses	Malformations		Weight (g)	
			Number	Percent	Mean	SD
0.00	25	297	1	0.337	0.972	0.098
0.75	24	276	26	9.42	0.877	0.104
1.50	22	229	89	38.865	0.764	0.107
3.00	23	226	129	57.08	0.704	0.124

- What is the joint effects of increasing dose on fetal weight (continuous) and on malformation status (yes/no)?

Another Example

- Depression clinical trial (2004)

Table 8.2 *Descriptive statistics for depression data.*

Group	Time	HAMD		CGI	
		Mean	SD	Mean	SD
Active	0	29.3	6.0	4.1	0.5
	1	23.3	5.9	3.6	0.6
	2	18.4	7.9	3.1	0.8
	3	15.6	8.8	2.9	1.0
	4	14.4	8.5	2.7	1.0
	5	10.6	6.0	2.6	0.9
	6	8.9	6.6	2.0	1.0
Control	0	31.1	6.3	4.3	0.7
	1	26.1	6.9	3.9	0.7
	2	20.9	7.6	3.5	0.8
	3	19.8	7.9	3.5	1.0
	4	18.6	8.6	3.4	1.0
	5	14.3	8.4	2.9	0.9
	6	11.7	6.5	2.5	1.1

HAMD: Hamilton Depression Rating Score (0-90); CGI: Clinical Global Improvement Score (1-7).

Another Example

- Manitoba Health Utilizations
  - Population-based analysis
  - <5% of the population use health care services
  - They are responsible for >95% of all healthcare expenditures

Introduction (cont'd)

- The goal is on how to describe mixed data
  - Distribution of the mixed data
  - Multivariate methods for mixed data
    - Two-sample test
    - Discrimination/classification
    - Distance methods with mixed data
- Another goal is on how to describe correlations/associations/ among mixed data
  - Joint analysis of mixed outcomes

Introduction (cont'd)

- Traditional approaches
  - Transform the data to a normal distribution
- New methods developed
  - But statistical software and packages did not keep pace with these advances
- Eg. Engineering, finance, medicine and health

Issues

- Multivariate data of same data types
- Multivariate data of mixture data types
- Multi-level data
- non-standard correlated data
  - Pearson correlation
  - Polychoric correlation (ordinal)
  - Tetrachoric correlation (binary)
  - Biserial correlation (binary and continuous)
  - Poliserial correlation (ordinal and continuous)

Objectives

- What is the nature of relationships between measurements?
  - Different and/or the same subjects?
  - Over time?
  - Cross-sectionally?

Approaches

- Separate analysis for the variables
  - Deficient in many applications
- Joint analysis not straightforward
- Significant developments for the past 2 decades.
- Ideal approach
  - Specify a model for the joint distribution
  - Fit the model to the data
  - Use the parameter estimates to draw inferences

History of Mixed Data Methods

- Early ad-hoc approaches
  - Subject the discrete variables on some numerical scoring scheme – treat all as continuous
  - Subject the continuous variables to some grouping criteria to discretize – treat all as discrete
  - Analyze separately and synthesize the two sets of results.
- All involve some element of subjectivity, loss of information, ignore association in mixed variables and unsatisfactory in general

History (cont'd)

- Ideal general approaches
  - Marginal distribution of the discrete variables and conditional distribution of the continuous variables given the discrete variables
  - Marginal distribution of the continuous variables and conditional distribution of the discrete variables, given the continuous variables

History (cont'd)

Given mixed random variables: (X, Y)

X : discrete

Y: continuous

(1): $F(X, Y) = F(Y|X)P(X)$

(2): $F(X, Y) = P(X|Y)F(Y)$

History (cont'd)

- Conditional Gaussian Distribution (CGD)
  - Received much attention
  - Different multivariate normal distribution for each setting of the categorical variable values,
  - Categorical variables with an arbitrary marginal multinomial distribution
    - Whittaker (1990), Lauritzen and Wermuth (1989)

History (cont'd)

- Conditional Gaussian regression models
  - logistic conditional distribution for the binary variables given the continuous variables, multiplied by a marginal multivariate normal distribution –
    - Cox(1972),
    - Cox and Wermuth (1992) connected to probit-style and latent variable models

CGD - Early models

- General location model (GLOM)
  - Continuous+nominal
  - Continuous variables follow multivariate normal
  - Discrete variables decide the location
- Conditional Grouped continuous model (CGCM)
  - Continuous+ordinal
  - Transform the ordinal variables into continuous variables - GCM
  - Distribution of the latent variable is conditional given the continuous data

CGD – GLOM

- Moustafa (1957) and Olkin and Tate (1961) were the first to consider the full CGD model.
  - Multi-way tables
  - Binary and continuous data
  - Canonical correlations between binary and continuous variables
  - Connection between canonical correlations and the state means.

CGD-GLOM

Given mixed random variables: (X, Y)

X : nominal

Y: continuous

$F(X, Y) = \Phi\left(\mu_X, \Xi_X\right)P(X)$

$\Phi\left(\mu_X, \Xi_X\right)$ multivariate normal with

mean: μ_X

variance matrix: Ξ_X

CGD-GLOM

- Developed as a device for hypothesis testing of independence and conditional independence
- The joint prob density of observing state *s:*

$$\pi_s(2\pi)^{-C/2}|\boldsymbol{\Sigma}_s|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_s)^{\top}\boldsymbol{\Sigma}_s^{-1}(\mathbf{y}-\boldsymbol{\mu}_s)\right\}.$$

- where *s* is the product of all possible patterns of discrete response.

CGD - GCM

- GCM considers multivariate normal distribution for a latent variable (Y*~NQ(0,R*)) underlying the Q category ordinal variable (Z) and partition the space of the latent variable into non-overlapping intervals ; Pearson (1904).
- R* are polychoric correlations.
- CGCM, assume a joint MVN with continuous variables, resulting in Polyserial correlations between the discrete and continuous variables

CGD-GCM

Given mixed random variables: (Y, Z)

Y: continuous

Z: ordinal$\rightarrow Z^*$: continuous

$F(Y, Z^*) = \Phi(\mu, \Xi)$

$\Phi(\mu, \Xi)$ multivariate normal with

mean: μ

variance matrix: Ξ

GMDM

- General mixed-data model (GMDM)
  - Continuous+nominal+ordinal
  - hybrid of GLOM and CGCM
  - Two components;
    - GLOM for nominal and continuous
    - CGCM for ordinal and continuous, given the nominal data

$$f_{\mathbf{X},\mathbf{Y},\mathbf{Z}}(\mathbf{x}_{(s)},\mathbf{y},\boldsymbol{\ell}) = \pi_s \phi_C(\mathbf{y}-\boldsymbol{\mu}_s|\boldsymbol{\Sigma}) \int_{S(s,\mathbf{y},\boldsymbol{\ell})} \phi_Q(\mathbf{v}|\mathbf{R})d\mathbf{v}$$

  - Nominal based on GLOM
  - Ordinal based on CGCM
  - De Leon and Carriere (2007)

CGD-GMDM

Given mixed random variables: (X, Y, Z)

X : nominal

Y: continuous

Z: ordinal$\rightarrow Z^*$: continuous

$F(X, Y, Z^*) = \Phi(\mu_X, \Xi_X)P(X)$

$\Phi(\mu_X, \Xi_X)$ multivariate normal with

mean: μ_X

variance matrix: Ξ_X

CGD

- General location model (GLOM)
  - Continuous+nominal
  - Continuous variables follow multivariate normal
  - Discrete variables decide the location
- Conditional Grouped continuous model (CGCM)
  - Continuous+ordinal
  - Transform the discrete variables into continuous variables
- General mixed-data model (GMDM)
  - Continuous+nominal+ordinal
  - Combination of GLOM and CGCM

Issues with factorization

- Uses structural approach, classifying outcomes to decide the direction of conditioning, inducing hierarchy
  - Condition – intermediate outcomes
  - Conditioned – primary responses
- Shortcomings;
  - Not invariant to the direction of conditioning
  - Models uncomparable,
  - Different interpretations for parameters.
  - Very different inferences, esp. in associations.

Indirect approaches

- Basic idea
  - Use random effects to build in correlation
  - Treat outcomes symmetrically
  - Flexibility in accounting for different measurement levels
  - Delineating of data settings.
- Shortcomings
  - Correlations may be restricted to lie within narrow ranges
  - Computational difficulties may arise in high dimensional problems

Indirect Approaches

- GLMM (Faes, 2008, 2009) in longitudinal setting and also in high dimensional cases
- GLLAMM (Rabe-Hesketh, She&Lee 2000, 2001) for latent and mixed
- Bayesian (Wagner and Tuchler, 2010, Daniels and Normand, 2006)

Copula

- Construct joint models for X and Y, where the relevant joint distribution is difficult.
- Studied for mixed continuous and ordinal data.
- Given random effect B,

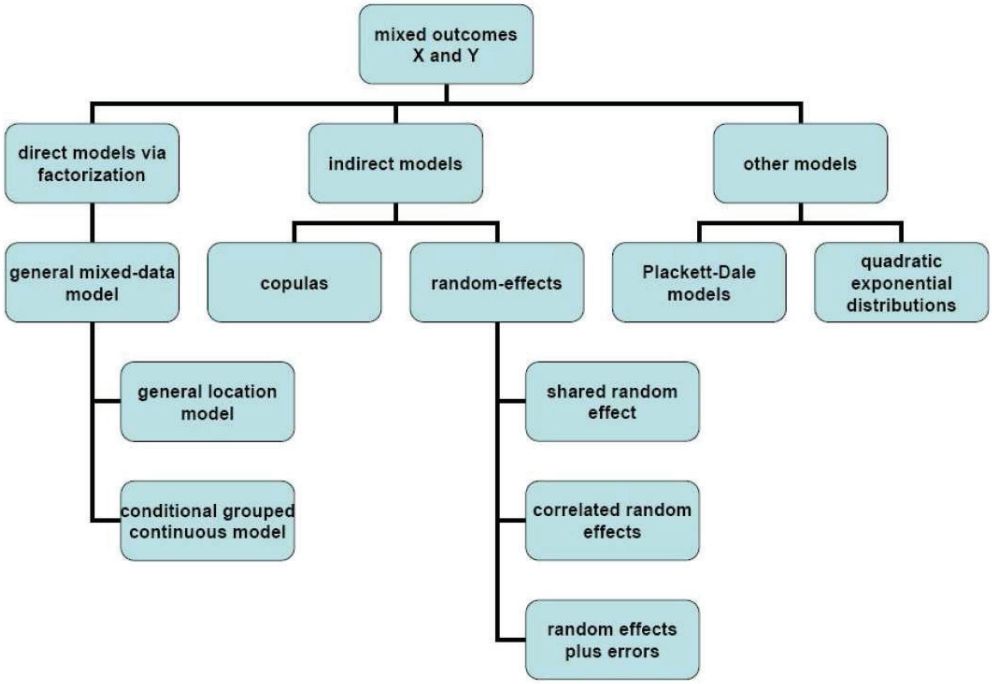
$$F_{x_1,\ldots,x_p}(x_1,\ldots,x_p) = C\{F_{x_1}(x_1),\ldots,F_{x_p}(x_p)\}$$

- where the p-dimensional Gaussian copula is

$$C(u_1,\ldots,u_p) = \Phi_p\{\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_p);\tilde{R}\}$$

- Discrete data applications are rare

Joint Analysis of Mixed Outcomes



GLOM and Copula

- 1. Transform ordinal data Y to be continuous and normal, say X^*
  - Then we have transformed the data into new mixed data: two continuous + one nominal variables
  - Generally speaking, the marginal distributions of $f(X)$, $f(X^*)$ and $P(Z)$ are easy to obtain, given the data.

GLOM and Copula

- 2. At each level of Z, i. e., Z_1 and Z_2, respectively, fit a joint distribution of X, and X^* , using copula method with marginal distribution of $f(X)$ and $f(X^*)$

$$F_{X,X^*}\left(X, X^*\right) = C\left(F_X(X), F_{X^*}\left(X^*\right)\right)$$

GLOM and Copula

- For Gaussian copula, we have

$$C\left[\mu(z_i), \mu^*(z_i)\right] = \Phi\left\{\phi^{-1}\left[\mu(z_i)\right], \phi^{-1}\left[\mu^*(z_i)\right]; \Xi(z_i)\right\}$$

- All are functions of Z=z_i, the value of the nominal variable Z

Estimation

- Highly sparse data
- High dimensional integration
  - Likelihood approaches
    - Pseudo, composite, pairwise LR
  - Data augmentation
  - GEE
  - Numerical integration,
  - MCMC
  - Bayesian

Estimation

- Joint estimation: MLE with parameters

$$\mu(z_i), \mu^*(z_i), \Xi(z_i)$$

- Marginal estimation using margins with IFM:

$$F_X(X), F_{X^*}(X^*)$$

Estimation (cont'd)

- MLE
  - Straightforward and its efficiency and consistency are well proven
  - but computationally expensive, especially with many variables/parameters
- IFM (Joe and Xu, 1996)
  - Computationally straightforward
  - May lack sound theoretical support and testing of its efficiency
  - Consistency?

Data Example

- Depression clinical trial (2004)

Table 8.2 *Descriptive statistics for depression data.*

Group	Time	HAMD		CGI	
		Mean	SD	Mean	SD
Active	0	29.3	6.0	4.1	0.5
	1	23.3	5.9	3.6	0.6
	2	18.4	7.9	3.1	0.8
	3	15.6	8.8	2.9	1.0
	4	14.4	8.5	2.7	1.0
	5	10.6	6.0	2.6	0.9
	6	8.9	6.6	2.0	1.0
Control	0	31.1	6.3	4.3	0.7
	1	26.1	6.9	3.9	0.7
	2	20.9	7.6	3.5	0.8
	3	19.8	7.9	3.5	1.0
	4	18.6	8.6	3.4	1.0
	5	14.3	8.4	2.9	0.9
	6	11.7	6.5	2.5	1.1

HAMD: Hamilton Depression Rating Score (0-90); CGI: Clinical Global Improvement Score (1-7).

Depression Clinical Trial

- The model:

$$
\begin{aligned}
Y_{ij1} &= \beta_{10} + \beta_{11} I_i + \beta_{12} j + \beta_{13} I_i j + \beta_{14} j^2 + B_{i11} + B_{i12} j + \varepsilon_{ij1}, \\
Y^*_{ij2} &= \beta_{20} I_i + \beta_{21} j + \beta_{23} I_i j + \beta_{24} j^2 + B_{i21} + B_{i22} j + \varepsilon_{ij2},
\end{aligned}
$$

where

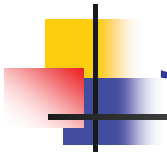
$$
\begin{pmatrix} B_{i11} \\ B_{i12} \\ B_{i21} \\ B_{i22} \end{pmatrix} \overset{iid}{\sim} N_4(\mathbf{0}, \mathbf{\Sigma}_b) \quad \text{and} \quad \begin{pmatrix} \varepsilon_{ij1} \\ \varepsilon_{ij2} \end{pmatrix} \overset{iid}{\sim} N_2(\mathbf{0}, \mathbf{\Sigma}_e).
$$

Depression Clinical Trial

- Joint vs separate modeling

Table 8.4 *MLEs (approximate) of regression parameters, error variance, and covariance, and their large-sample SEs for depression data.*

Parameter	Separate analyses				Joint model	
	HAMD		CGI			
	Est	SE	Est	SE	Est	SE
HAMD						
β_{10}: intercept	3.08	0.13	—	—	3.08	0.13
β_{11}: treatment	–0.21	0.17	—	—	–0.21	0.17
β_{12}: time	–0.50	0.05	—	—	–0.49	0.05
β_{13}: treatment-by-time	–0.06	0.05	—	—	–0.06	0.07
β_{14}: quadratic time	0.04	0.01	—	—	0.03	0.01
$\sigma^2_{\varepsilon 1}$: error variance	0.15	0.01	—	—	0.15	0.01
$\sigma_{\varepsilon 12}$: error covariance	—	—	—	—	0.30	0.03
CGI						
τ_1: threshold 1	—	—	–6.60	0.61	–6.07	0.55
τ_2: threshold 2	—	—	–3.79	0.42	–4.03	0.43
τ_3: threshold 3	—	—	–1.81	0.34	–2.01	0.35
τ_4: threshold 4	—	—	0.87	0.32	0.90	0.31
β_{21}: treatment	—	—	–0.41	0.37	–0.52	0.40
β_{22}: time	—	—	–0.65	0.19	–0.80	0.18
β_{23}: treatment-by-time	—	—	–0.28	0.18	–0.22	0.15
β_{24}: quadratic time	—	—	0.02	0.02	0.05	0.02



Joint modelling

- Efficiency gain is relatively small when shared covariates are involved,
- But the efficiency gain is high in general and high with a large proportion of missing data or with an extreme case of binary data (near 0 or 1)
- Main gain in understanding the association of the outcomes

Acknowledgements

- Natural Sciences and Engineering Research Council of Canada
- Alberta Heritage Foundation for Medical Research