Modeling Data from Cluster Randomized Trials in Presence of Measurement Error Induced by Double Clustering

Maria Esther Perez Trejo
Department of Epidemiology, Biostatistics and Occupational Health
McGill University

April 8th 2011

Outline

- Background
  - Cluster RCT: advantages/disadvantages.
  - Double clustering situation.
  - Study that illustrates double clustering: PROBIT.
  - Strategies to minimize double clustering

- Modeling double clustering
  - Why not traditional cluster RCT analysis tools?
  - Bayesian models with missing data for continuous outcomes.

- Analysis of PROBIT data

- Simulation study
  - MCMC approach WinBUGS-R
  - Simulation scenarios
  - Simulation results

- Conclusions and Future work

Background

- **Advantages of cluster RCTs**

  - Benefits inherent to randomized treatment allocation while reducing potential contamination of treatment effects due to existence of subjects in close proximity.

  - More convenient than individual RCT when group dynamics make more feasible to change practices/behaviors within a group than among individuals.

- **Disadvantages of cluster RCTs**

  - Reduced statistical power for treatment effect estimation due to within cluster correlation of outcomes.

    * within cluster correlation measured through the ICC.

    * this problem can be addressed at the design stage by increasing the sample size according to the magnitude of ICC.

  - When the number of clusters is small, imbalance can occurr in potential confounders that vary across clusters.

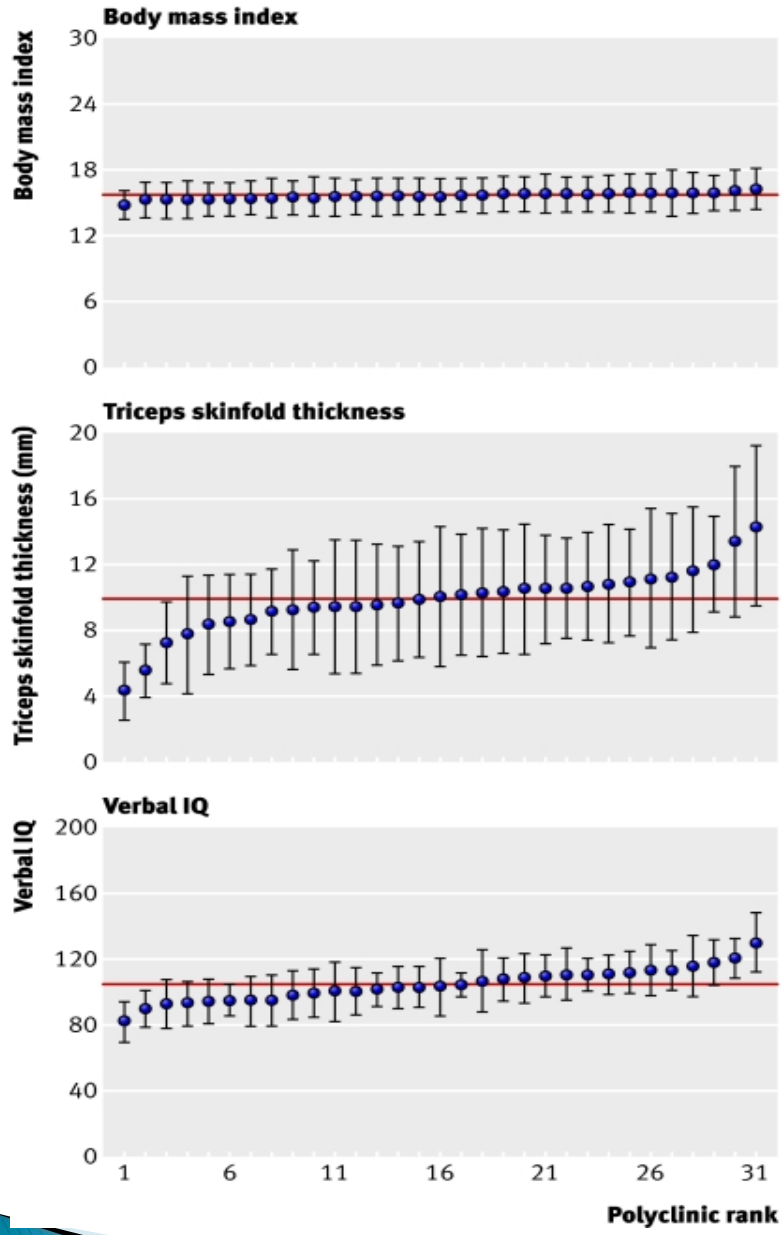
    * problem addressed through statistical adjustment.

Double Clustering

- When the measurement of the outcome is naturally clustered, and that clustering occurs within the same groups used as units for cluster randomization.
- A cluster is now defined as the group of subjects whose outcome is measured by the same observer.
- Difference in treatment effects due to cluster measurement cannot be separated from differences caused by inherent clustering.
- Effect: to increase degree of association between individuals within the same cluster → decrease detection rate of treatment effects.
- It cannot be addressed at a design level, neither through statistical adjustment.

PROBIT (Promotion of Breastfeeding Intervention Trial)

- Units of randomization: hospitals.
- 17,046 children originally randomized among 31 hospitals in Belarus.
- Treatment: breastfeeding promotion intervention/standard care.
- Outcomes of interest: body mass index (BMI), triceps skinfold thickness (TST), verbal IQ (IQ).
- A different observer/equipment per site took these measurements $\Rightarrow$ double clustering is present!



BMI: digital read out weight scale the least susceptible to between hospitals differences ⇒ BMI does not vary considerably by cluster.

TST and IQ: ranges are too variable to be explained by true geographical differences ⇒ differences likely to reflect systematic measurement differences among the 31 clusters.

The ICC's reflect this situation:

- ICC for BMI = 0.03
- ICC for TST = 0.18
- ICC for IQ = 0.31

Strategies to minimize double clustering

- To randomly allocate observers across clusters.
  * It might not be feasible due to geographical limitations.

- To use a single individual to assess the outcome in all clusters → **audited measures in each cluster.**
  * Complicated by large number of participants or geographical dispersion.
  ****Statistical model of this strategy presented here!**

- To standardize measurement methods/training of observers.

- Pilot study to identify difficulties in outcome measurements.

- Why modeling double clustering is important?
  - It is very likely to occurr in cluster RCTs.
  - It could partially explain some negative/inaccurate results of some cluster RCTs carried out in the past.

Modeling Double Clustering

- Traditional modeling for cluster RCTs continuous outcomes

$$Y_{ij}^{(1)} = \beta_0 + \beta_1 T_i + b_i + \varepsilon_{ij}^{(1)} \qquad (1)$$

- $Y_{ij}^{(1)}$ = outcome for the *j*th subject in the *i*th cluster.
- T_i = treatment allocated to cluster *i* (*T* = 1 for intervention, *T* = 0 for control).
- $b_i \sim N(0,\sigma_b)$ and $\varepsilon_{ij}^{(1)} \sim N(0, \sigma_1)$ are the error terms at cluster and individual level, respectively.
- Independence of errors assumed.
- ICC given by

$$ICC = \frac{\sigma_b^2}{\sigma_1^2 + \sigma_b^2}$$

- Usual methodology: linear/generalized mixed models.
  -clustered accounted for through random intercepts estimation.
  - Individual observations within each cluster considered as repeated measures.

- However, in presence of double clustering the model becomes

$$Y_{ij}^{(2)} = \beta_0 + \beta_1 T_i + b_i + d_i + \varepsilon_{ij}^{(2)} \qquad (2)$$

- $d_i \sim N(0,\sigma_d)$ is the random effect due to measurement error (cluster measurement). Independent of random terms b_i and $\varepsilon_{ij}^{(2)} \sim N(0, \sigma_2)$.
- ICC given by

$$ICC_{dc} = \frac{\sigma_b^2 + \sigma_d^2}{\sigma_2^2 + \sigma_b^2 + \sigma_d^2}$$

- If linear/generalized models are used to estimate treatment effects under model (2)

  - estimates will be inaccurate due to increased total variance.

  - not possible to separate variance due to clustering (σ_b) from variance due to measurement error (σ_d).

Proposed model: Audited data

- Observed data under double clustering come from model (2).
- n_i individuals per cluster $n_i \leq S_i$ (S_i denotes cluster size) are measured by a single auditor for all clusters.
- Audited outcomes come from model (1) ($\sigma_d = 0$).

$\Rightarrow$

- Bayesian inference in presence of missing data.
- n_i individuals per cluster have both $Y_{ij}^{(1)}$ (audited) and $Y_{ij}^{(2)}$ outcomes.
- S_i - n_i subjects per cluster have only $Y_{ij}^{(2)}$ outcome; for these individuals $Y_{ij}^{(1)}$ is a missing observation.

- Model fit in WinBUGS
- 2 separated error terms for models (1) and (2)

  - pilot simulation study (in R-WinBUGS) produced more accurate cluster and error term variances than under a single error term scenario.

  - feasible this situation to happen in real life.

Formal model for continuous outcomes

- MCMC approach to estimate parameters of interest from a suitable posterior distribution.

- Define vector of parameters

$\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma_1, \sigma_2, \sigma_b, \sigma_d, b_i, d_i, i = 1,\ldots,n_c$ where n_c is the number of clusters).

- Main interest lies in estimating the treatment effect (β_1) and variances for cluster and measurement error (σ_b and σ_d, respectively).

Distribution of outcomes for one cluster (i^{th} cluster)

- Assume all clusters have the same size S.
- Vector $\boldsymbol{Y}_i$ of (2 S) entries of observations can be written as:

$$\boldsymbol{Y}_i = [Y_{i1}^{(1)}, Y_{i2}^{(1)},\ldots,Y_{is}^{(1)}, Y_{i1}^{(2)}, Y_{i2}^{(2)},\ldots,Y_{is}^{(2)}]^{\intercal}$$

where terms $Y_{ij}^{(2)}$ are the S observed outcomes from model (2) -double clustering, and terms $Y_{ij}^{(1)}$ are the n_i observed (audited) outcomes and the S- n_i missing observations from model (1).

- It is assumed that

$$\boldsymbol{Y}_i \sim \text{MVN}_{(2s\ \ 1)}(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$$

where

$$\boldsymbol{\mu}_i = [\beta_0 + \beta_1 T_i ,\ldots, \beta_0 + \beta_1 T_i\ ]^{\intercal},$$

and $\boldsymbol{\Sigma}_i$ is the (2s 2s) matrix for the variance-covariance structure given by

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} D_i & P_i & \dots & P_i \\ P_i & D_i & \dots & P_i \\ \dots & \dots & \dots & \dots \\ P_i & P_i & \dots & D_i \end{bmatrix}$$

- *S* (2 2)-matrices D_i showing the covariance structure between outcomes of the same subject under both models.
- Two observations of the same subject

$$\boldsymbol{Y}_{ij} = [\boldsymbol{Y}_{ij}^{(1)}, \boldsymbol{Y}_{ij}^{(2)}]^{\mathsf{T}}$$

assumed bivariate normal with vector mean

$$\boldsymbol{\mu} = [\beta_0 + \beta_1 T_i , \beta_0 + \beta_1 T_i]^{\mathsf{T}}$$

and variance-covariance structure

$$D_i = E\left[\left(Y_{ij} - \boldsymbol{\mu}\right)^{\mathsf{T}}\left(Y_{ij} - \boldsymbol{\mu}\right)\right] = \begin{bmatrix} \sigma_b^2 + \sigma_d^2 + \sigma_1^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_2^2 \end{bmatrix}$$

- The (2 2) matrices P_i out of the diagonal show the covariance structure between outcomes of two different individuals.

- P_i matrices given by

$$P_i = E\left[\left(Y_{ij} - \mu\right)^{\mathrm{T}}\left(Y_{ij+1} - \mu\right)\right] = \begin{bmatrix} \sigma_b^2 + \sigma_d^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 \end{bmatrix}$$

- Joint pdf for the observations of a cluster is

$$p(Y_i|\theta) = \frac{1}{(2\pi)^{2s/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(Y_i - \mu_i)^T \Sigma_i (Y_i - \mu_i)\right\}$$

Joint distribution of observations for all clusters

- Alternative expression of observations in cluster i in terms of the vectors of observed outcomes $\boldsymbol{Y}_{i\text{obs}}$, and missing outcomes $\boldsymbol{Y}_{i\text{mis}}$, $\mathbf{Y}_i = [\boldsymbol{Y}_{i\text{obs}}, \boldsymbol{Y}_{i\text{mis}}]$.

- Vector containing the $(2S \;\; n_c)$ observations from the n_c clusters is $\boldsymbol{Y} = (\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_{nc}) = (\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}})$, where $\boldsymbol{Y}_{\text{obs}} = (\boldsymbol{Y}_{1\text{obs}},\ldots,\boldsymbol{Y}_{nc\text{obs}})$, and $\boldsymbol{Y}_{\text{mis}} = (\boldsymbol{Y}_{1\text{mis}},\ldots,\boldsymbol{Y}_{nc\text{mis}})$.

- Independence between observations from different clusters is assumed. Therefore, joint distribution of $\boldsymbol{Y}$ given parameters $\boldsymbol{\theta}$ is

$$p\left(\boldsymbol{Y}_{obs},\boldsymbol{Y}_{mis}\,\middle|\,\boldsymbol{\theta}\right) = \prod_{i=1}^{nc} \frac{1}{(2\pi)^{2s/2}|\boldsymbol{\Sigma}_i|^{1/2}} exp\left\{-\frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\right\} \qquad (3)$$

Prior distributions

- Intercept and treatment effects: β_0 and $\beta_1 \sim N(0, 1e^{-06})$

- Diffuse priors for variation parameters:
  $1/\sigma_b^2, 1/\sigma_d^2, 1/\sigma_1^2, 1/\sigma_2^2 \sim$ Gamma(0.001,0.001)

- For cluster and measurement error terms:
  $b_i|\sigma_b \sim N(0,\sigma_b)$ and $d_i|\sigma_d \sim N(0,\sigma_d)$ for $i = 1,\ldots,n_c$

- Independence between clusters is assumed, then

$$p(b_1,\ldots,b_{nc}|\sigma_b) = \prod_{i=1}^{nc} p(b_i|\sigma_b) \quad \text{and} \quad p(d_1,\ldots,d|\sigma_b) = \prod_{i=1}^{nc} p(d_i|\sigma_b)$$

- Joint priors for cluster and measurement error terms:

$$p(b_1,..,b_{nc},\sigma_b) = p(b_1,..,b_{nc}|\,\sigma_b)p(\sigma_b) \quad \text{and} \quad p(d_1,..,d_{nc},\sigma_d) = p(d_1,..,d_{nc}|\,\sigma_d)\,p(\sigma_d)$$

- Joint prior for $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}) = p(\beta_0)\,p(\beta_1)\,p(\sigma_1)\,p(\sigma_2)\,p(b_1,..,b_{nc},\sigma_b)\,p(d_1,..,d_{nc},\sigma_d) \qquad (4)$$

Posterior distribution for MCMC

- Since missing observations are unknown, multiple imputation is carried out → MCMC sampling from the posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{Y}_{\text{mis}}$ is conducted.

- MCMC sample from

$$p(\boldsymbol{Y}_{mis}, \boldsymbol{\theta} | \boldsymbol{Y}_{obs}) = \frac{p(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{Y}_{obs})}$$

  where $p(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}} | \boldsymbol{\theta})$ given by equation (3) and $p(\boldsymbol{\theta})$ given by equation (4), and

$$p(\boldsymbol{Y}_{obs}) = \int_{\theta} \int_{Y_{mis}} p(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{Y}_{\text{mis}} \, d\boldsymbol{\theta}$$

- A GS-MH algorithm to generate a MCMC for the parameters of interest only requires to know that the posterior distribution is

$$p(\boldsymbol{Y}_{\text{mis}}, \boldsymbol{\theta} | \boldsymbol{Y}_{\text{obs}}) \propto p(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}} | \boldsymbol{\theta}) \, p(\boldsymbol{\theta}) \qquad (5)$$

Analysis PROBIT data (IQ)

- Without accounting for double clustering

Parameter	Posterior mean	Posterior sd	Posterior median	95% credible interval
β_0	107.4	2.3	107.5	(102.1, 112.5)
β_1	-5.48	2.7	-5.16	(-11.3, -1.24)
σ_1^2	183.4	2.2	183.4	(178.9, 187.6)
σ_2^2	97.8	29.2	92.7	(54.6, 168.4)
σ_b^2	2.81	4.4	0.38	(0.002, 14.6)

- Accounting for double clustering

Parameter	Posterior mean	Posterior sd	Posterior median	95% credible interval
β_0	106.4	2.4	106	(102.9, 111.1)
β_1	-2.15	2.5	-1.76	(-6.8, 2.51)
σ_1^2	199.1	21.7	198.3	(161.5, 245.3)
σ_2^2	183.5	2.2	183.6	(179.3, 188)
σ_b^2	54.9	18.8	51.5	(28.7, 100.5)
σ_d^2	42.4	16.8	39.7	(19.03, 86.6)

- Treatment effect shifted towards the null. 95% CI length practically unchanged.
- ICC = 0.22, and ICC_{dc} = 0.34 → about half of the ICC is due to double clustering variability.

Simulation Study

- MCMC sampling from posterior distribution (5) → for β_0, β_1, σ_1^2, σ_2^2, σ_b^2, σ_d^2: estimate posterior mean, variance and 95% credibility interval.

- *Simulation setting*:

  - 100 replications of each scenario.
  - $\beta_0 = 100$, $\beta_1 = 5$ (treatment effect), $\sigma_b=\sigma_1=\sigma_2=1$.

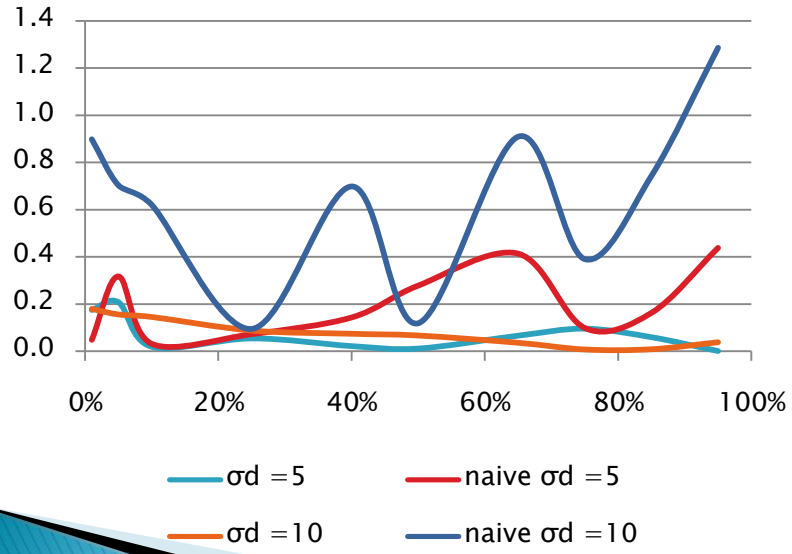
  - two scenarios for severity of double clustering: $\sigma_d= 5$ (moderate) and $\sigma_d= 10$ (strong).
  -10 scenarios for number of audited data: 1%, 5%, 10%, 25%, 40%, 50%, 65%, 75%, 85% and 95% of audited observations per cluster.

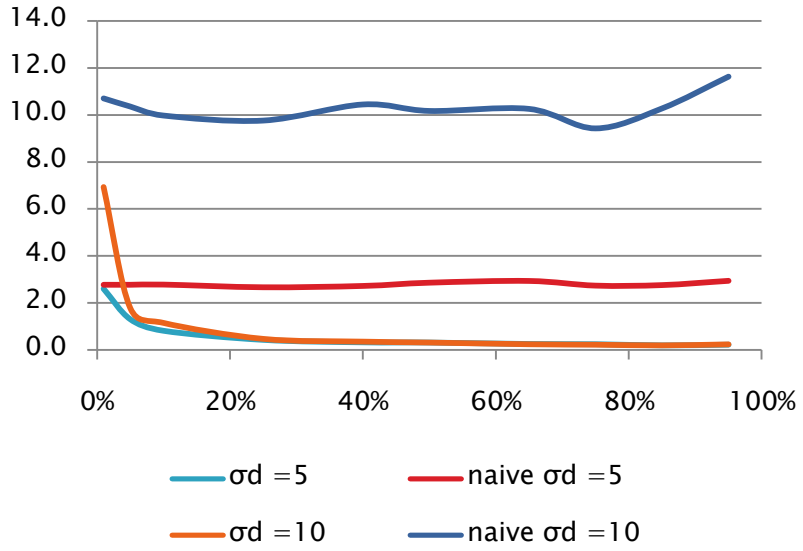
  - For each (% audited data, σ_d) combination posterior summary statistics obtained using **only** observed data in presence of double clustering from equation (2) → naive approach.

  -For each (% audited data, σ_d) scenario quality of estimator for each parameter evaluated through: bias, variance, MSE, relative efficiency (with respect to estimator under naive approach), coverage of credibility interval.

Treatment effect (β_1) –bias and MSE

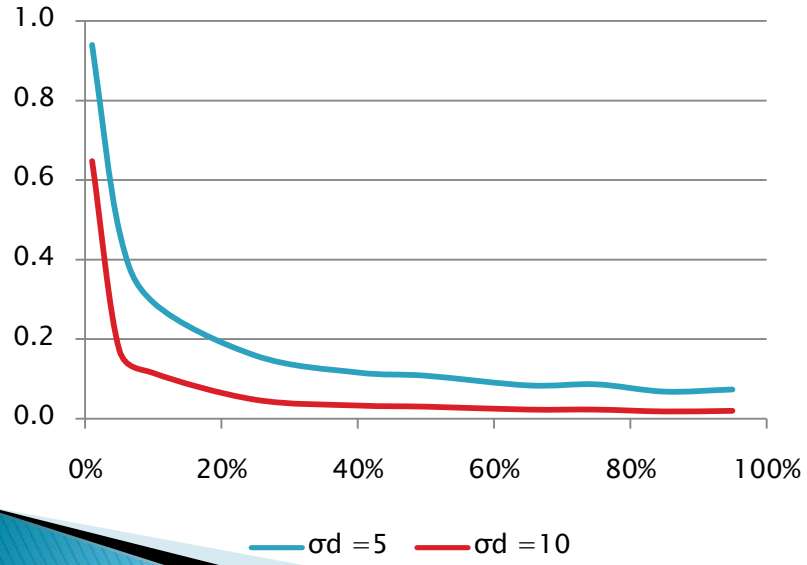
	Bias				MSE			
% of audited data	σ_d =5	naive σd =5	σ_d =10	naive σd =10	σ_d =5	naive σd =5	σ_d =10	naive σd =10
1%	0.175	0.048	0.179	0.898	2.604	2.773	6.930	10.704
5%	0.208	0.317	0.156	0.703	1.306	2.772	1.785	10.358
10%	0.019	0.031	0.145	0.620	0.812	2.781	1.146	9.972
25%	0.054	0.072	0.088	0.094	0.423	2.666	0.466	9.757
40%	0.022	0.143	0.074	0.698	0.317	2.722	0.349	10.444
50%	0.011	0.278	0.067	0.118	0.310	2.861	0.310	10.168
65%	0.066	0.414	0.036	0.909	0.247	2.933	0.237	10.256
75%	0.096	0.098	0.007	0.390	0.238	2.736	0.218	9.423
85%	0.059	0.162	0.008	0.743	0.189	2.759	0.189	10.276
95%	0.001	0.438	0.038	1.287	0.217	2.940	0.232	11.634

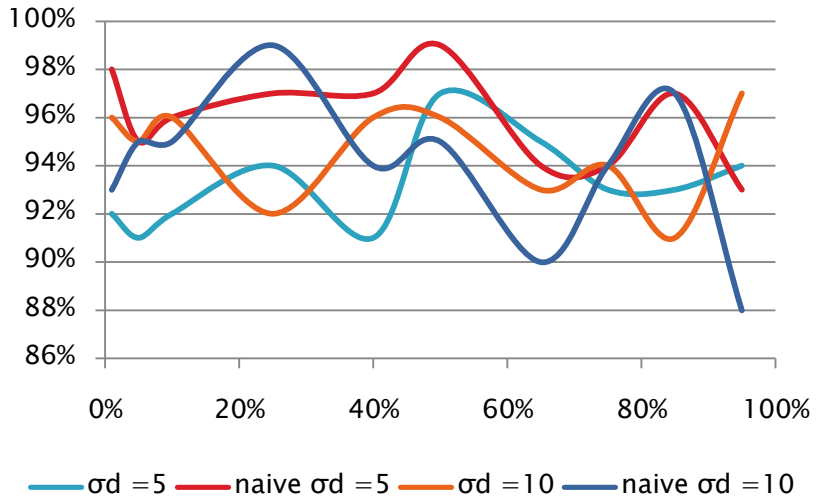




Treatment effect (β_1) –relative efficiency and coverage

	Relative efficiency		Coverage of CI			
% of audited data	σ_d =5	σ_d =10	σ_d =5	naive σd =5	σ_d =10	naive σd =10
1%	0.939	0.647	92%	98%	96%	93%
5%	0.471	0.172	91%	95%	95%	95%
10%	0.292	0.115	92%	96%	96%	95%
25%	0.159	0.048	94%	97%	92%	99%
40%	0.117	0.033	91%	97%	96%	94%
50%	0.108	0.030	97%	99%	96%	95%
65%	0.084	0.023	95%	94%	93%	90%
75%	0.087	0.023	93%	94%	94%	94%
85%	0.068	0.018	93%	97%	91%	97%
95%	0.074	0.020	94%	93%	97%	88%





Treatment effect (β_1)

- **Bias**

  - It tends to decrease as the % audited data increases for both scenarios of double clustering severity, although trend is not constant.

  - Degree of double cluster severity does not have impact on bias for estimated treatment effect (crossing of values of bias for some % audited data levels).

  - Bias when naive approach is used is considerably larger than under missing data approach for the σ_d= 10 case than for the σ_d= 5 scenario.

- **MSE**

  - Clear and steady decreasing trend as the % audited data increases for both scenarios of double cluster severity.

  - Values of MSE very similar at each % audited data level for both double clustering scenarios.

  - MSE under naive approach considerably larger when double clustering is severe than for a moderate level.

Treatment effect (β_1)

- **Relative efficiency**

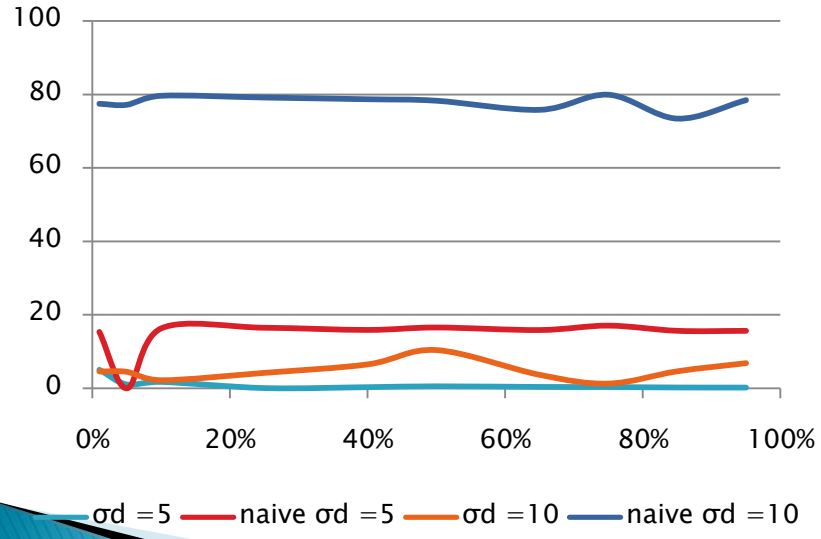
  - Steady decreasing trend as the % audited data increases.
  - Relative efficiency larger for moderate double clustering severity than under strong double clustering.

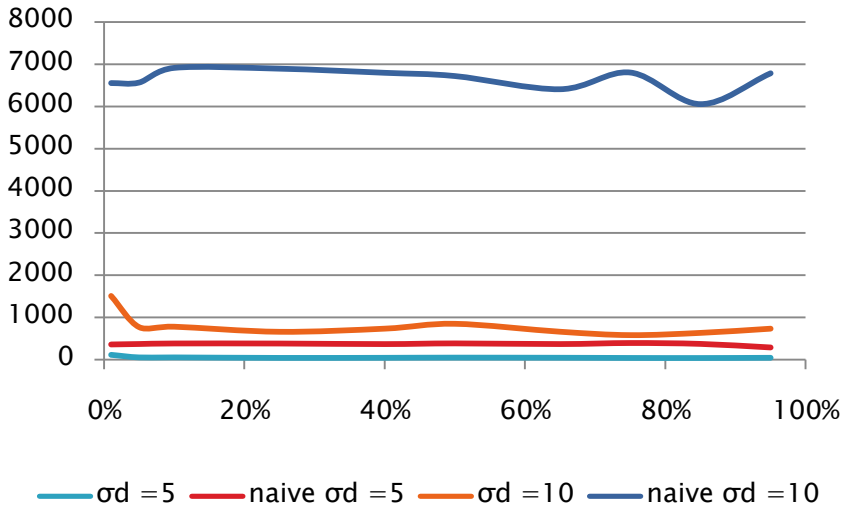
- **Coverage of CI**

  - Very close to theoretical level of 95% for all combinations of scenarios of double cluster severity and % audited data.
  - Also very close to theoretical level for both naive cases.

Measurement error variability (σ_d^2) – bias and MSE

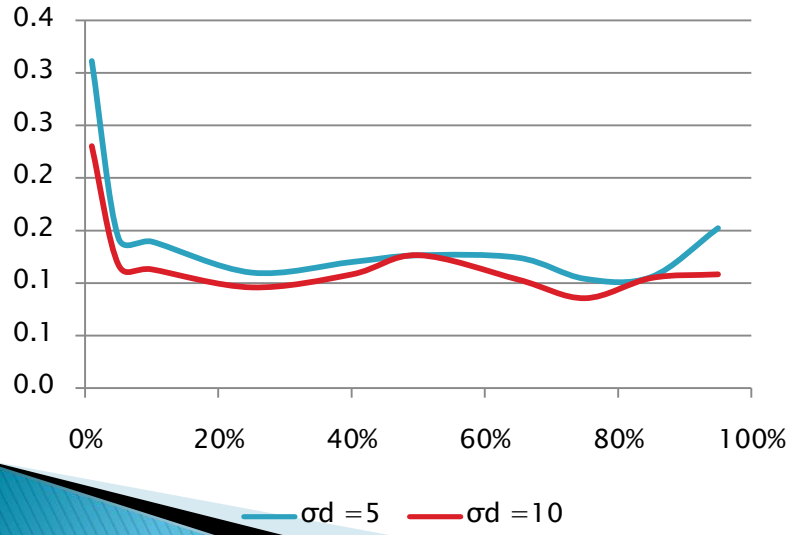
	Bias				MSE			
% of audited data	σ_d =5	naive σd =5	σ_d =10	naive σd =10	σ_d =5	naive σd =5	σ_d =10	naive σd =10
1%	5.060	15.393	4.627	77.439	111.94	359.85	1509.82	6555.9
5%	1.070	16.064	4.521	77.159	52.81	370.08	769.97	6567.9
10%	1.774	16.378	2.220	79.628	53.21	382.19	780.56	6914.3
25%	0.105	16.538	4.238	79.151	41.83	381.15	660.01	6895.0
40%	0.351	15.941	6.562	78.680	44.14	368.43	734.49	6797.6
50%	0.585	16.606	10.482	78.273	48.52	384.39	849.66	6720.8
65%	0.406	15.914	3.691	75.784	45.69	368.70	661.10	6410.3
75%	0.382	17.122	1.305	79.914	40.79	392.72	581.13	6804.0
85%	0.263	15.698	4.664	73.380	39.35	372.46	634.23	6058.4
95%	0.242	15.694	6.897	78.433	44.14	290.38	734.02	6786.7

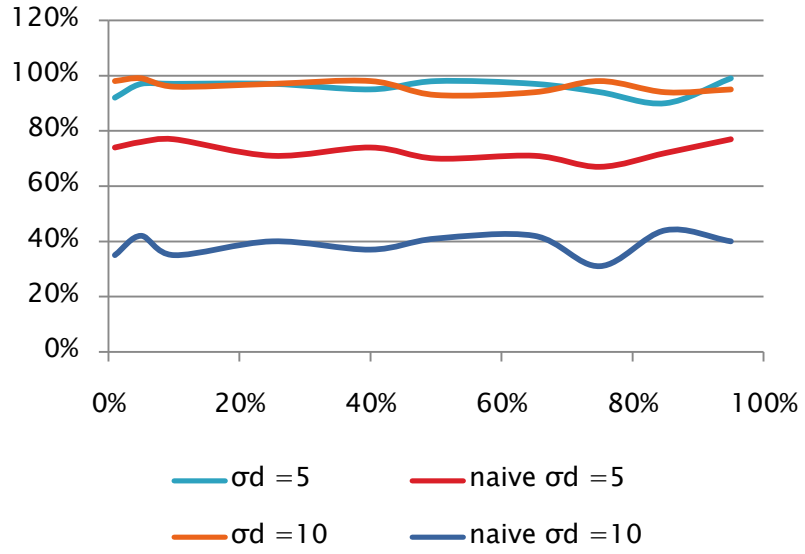




Measurement error variability (σ_d^2) – R.efficiency and coverage

	Relative efficiency		Coverage of CI			
% of audited data	σ_d =5	σ_d =10	σ_d =5	naive σd =5	σ_d =10	naive σd =10
1%	0.311	0.230	92%	74%	98%	35%
5%	0.143	0.117	97%	76%	99%	42%
10%	0.139	0.113	97%	77%	96%	35%
25%	0.110	0.096	97%	71%	97%	40%
40%	0.120	0.108	95%	74%	98%	37%
50%	0.126	0.126	98%	70%	93%	41%
65%	0.124	0.103	97%	71%	94%	42%
75%	0.104	0.085	94%	67%	98%	31%
85%	0.106	0.105	90%	72%	94%	44%
95%	0.152	0.108	99%	77%	95%	40%





Measurement error variability (σ_d^2)

- **Bias**

  - Decreasing trend as % audited data increases, when double cluster severity is moderate; however behavior is irregular as a function of % audited data for severe double clustering scenario.
  - For all % audited data, bias is smaller when double clustering is moderate than when it is severe.
  - Bias under naive approach remarkably higher than under presence of audited data, with difference being considerably more pronounced when double clustering is severe.

- **MSE**

  - Decreasing trend for small %'s audited data (up to 25%), then it fluctuates around 43 for scenario of moderate double clustering.
  - MSE for case of severe double clustering higher than when double clustering is moderate, for all cases of % audited data, although it does not follow any trend as function of % audited data.
  - MSE under naive approach considerably higher than when using audited data, with difference more pronounced when double clustering is severe.

Measurement error variability (σ_d^2)

- **Relative efficiency**

  - It does not show a considerable variability as % audited data changes, for both double cluster severity scenarios.
  - Very similar for both double clustering scenarios, for all levels of % audited data (around 0.15).

- **Coverage of CI**

  - Very close to theoretical level for all (% audited data, double clustering severity) scenarios.
  - It decreases considerably with respect to theoretical level when naive approach is used with levels around 75% for moderate double cluster severity, and around 40% for stronger double clustering scenario.

Other parameters

- Results for intercept very similar to those for treatment effect.

- Results for cluster variance (σ^2_b) similar to results for measurement error variance (σ^2_d).

- Individual variance under no double clustering (σ^2_1):

  -Bias very similar for both double cluster severity scenarios, without a decreasing trend as % audited data increases; however, bias under naive approach is smaller in both cases.

  - MSE shows decreasing trend as % audited data increases for both double clustering severity scenarios (MSE very similar for both σ_d cases and 25% and up audited data), and in both cases MSE under naive approach is smaller.

  - Coverage levels very close to theoretical level (95%) for all combinations of (% audited data, double clustering severity), as well as under naive approach.

Other parameters (cont.)

- Individual variance under double clustering (σ^2_2):

  - Bias for both double clustering severity scenarios shows irregular trend as % audited data changes; bias similar in both cases and they do not differ considerably from those under naive approach.
  - MSE very constant (around 5) for all (% audited data, double clustering severity) combinations, as well as for naive approach.
  - Coverage levels very close to theoretical level of 95% for all all (% audited data, double clustering severity) combinations and naive case.

- Individual-level variances are less susceptible to double clustering than cluster and measurement error variances.

Conclusions

- Estimated treatment effect is affected negatively by the presence of double clustering. The accuracy and precision of the estimated effect improve when at least 5% of the outcomes are audited by a single observer.

- Estimated values of both variability paramenters (cluster and measurment error) are considerably affected in their accuracy and precision when no data are audited. These parameters are more sensitive to an increase of the double cluster severity than the treatment effect.

- Inclusion and modeling of data audited by a single observer in all clusters successfully account for double clustering, therefore improving the accuracy and precision of estimated treatment effects as well as of cluster and measurement error variance parameters.

The Sequential Conditional Expectation-Maximization (SCEM) algorithm

Michael Regier, Erica E. Moodie
McGill University,
Department of Epidemiology, Biostatistics,
and Occupational Health

CANNeCTIN

8 April 2011

Outline

The context: Marginal structural model (MSM)

- The methodological context is that we are interested in obtaining the marginal effect of treatment on an outcome when
  1. A set of covariates, X, confound the treatment, A, and the outcome, Y and
  2. Not all treatments are observable (i.e. there is a censoring mechanism).

Simplifying the MSM context

- We will consider the point-treatment context where at least one of the confounders is measured with error.
- It is a first step towards
  - Understanding the inherent complexities of the problem,
  - Identifying possible methodologies for unbiased estimation in the presence of measurement error, and
  - Extending the methods to more complex systems.

Point-treatment using counterfactuals

Under the counterfactual framework,

- Let a be a possible value of A, and
- Y_a denote the potential response we would expect to observe if the subject followed treatment a.
  - $Y_{a=1}$ denote a subject's outcome if treated, and
  - $Y_{a=0}$ if untreated.
- For a continuous outcome, we want to estimate the marginal effect of treatment
- We use the marginal structural model $\mathrm{E}[Y_a] = g(A : \beta)$ where β parameterizes the model.
- For the point treatment scenario we consider $\mathrm{E}[Y_a] = \beta_0 + \beta_1 A$.

Point-treatment details

- Under the assumptions of consistency, exchangeability (Robins 1999), positivity (Hernan and Robins 2006), and time ordering where exposure precedes outcome (Mortimer 2005), we can obtain unbiased estimates of β.
- The estimates are obtained using a weighted M-estimator

$$\sum_{i=1}^{n} [W_i^{-1}]^{\mathsf{T}} (Y_i - \beta_0 - \beta_1 A_i)$$

- Through the creation of a psuedo-population, the weighting breaks the confounding relationships between the confounding prognostic factors and the treatment.
  - $Y_a \coprod A|X$

Point-treatment stabilized weight

For the point-treatment scenario with censored treatments (Hernan et al. 2001), the stabilized weight for the i^{th} individual is

$$sw_i = \frac{p(A_i = a_i | C_i = 0) p(C_i = 0)}{p(A_i = a_i | C_i = 0, X_i = x_i) p(C_i = 0 | X_i = x_i)}$$

The focus of the investigation: The denominator

The key problem lies in the denominator,

$$p(A_i = a_i | C_i = 0, X_i = x_i) p(C_i = 0 | X_i = x_i)$$

- If a confounder X is unobservable but a proxy is used X^* then we are using

$$p(A_i = a_i | C_i = 0, X_i^* = x_i^*) p(C_i = 0 | X_i^* = x_i^*)$$

Measurement error

Two general types

- Classical measurement error models
  - The conditional distribution of X^* given X is modelled
- Regression calibration models
  - The conditional distribution of X given X^* is modelled
  - Berkson error models

Focus: Classical additive error model

The classical unbiased additive error model for the i^{th} subjects is

$$X_i^* = X_i + \epsilon$$

where

- X_i is the unobserved variable,
- $E(\epsilon|X_i) = 0$, and
- $Var(\epsilon|X_i) = \tau^2$.

Implications of measurement error

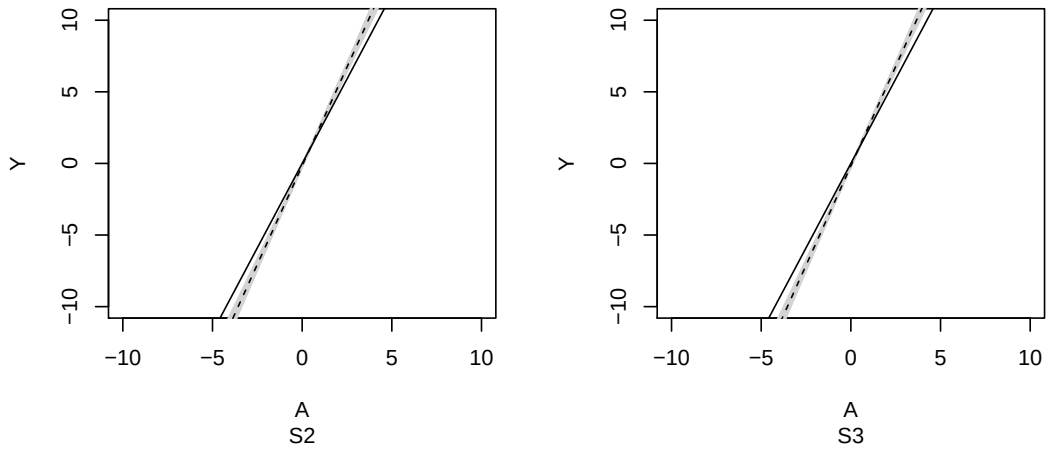
- Attenuation occurs in linear regression models when covariates are mismeasured.
- Gustafson shows that attenuation can be expected for parameters associated with mismeasured covariates in logistic regression.
- For both ordinary least squares and logistic regression attenuation is enhanced as the correlation amongst covariates strengthens (Gustafson 2004).

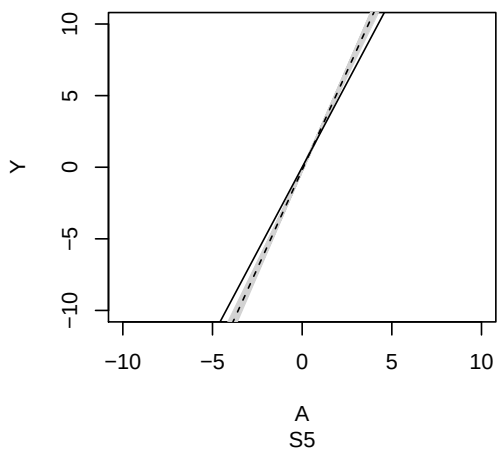
Effect of using proxy confounders

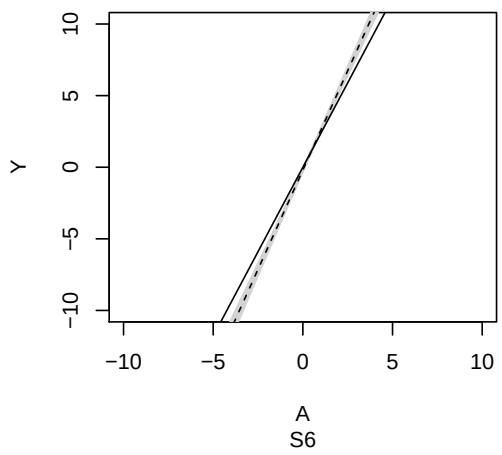
There are *four* effects of using proxy confounders in the denominator of the MSM weights on the parameter of interest, β_1:

1. Little effect,
2. Attenuation,
3. Augmentation, and
4. Sign reversal.

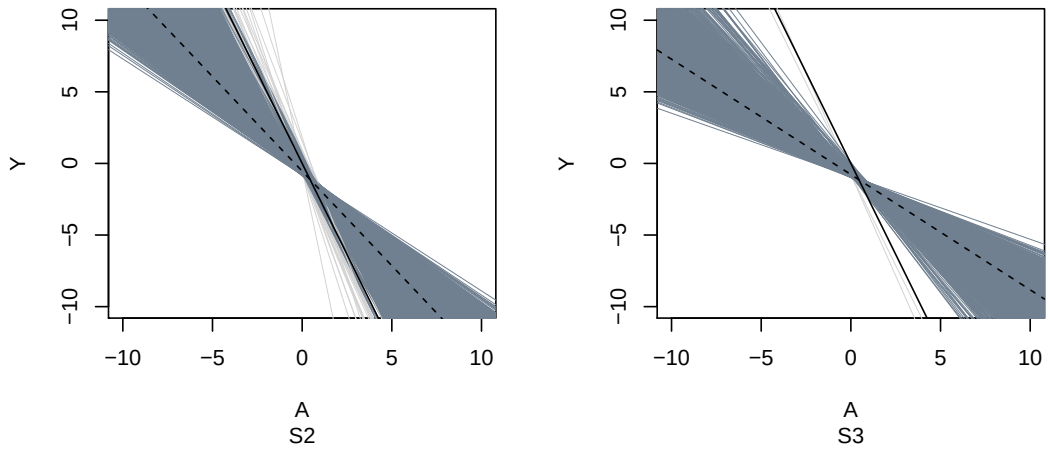
Little effect



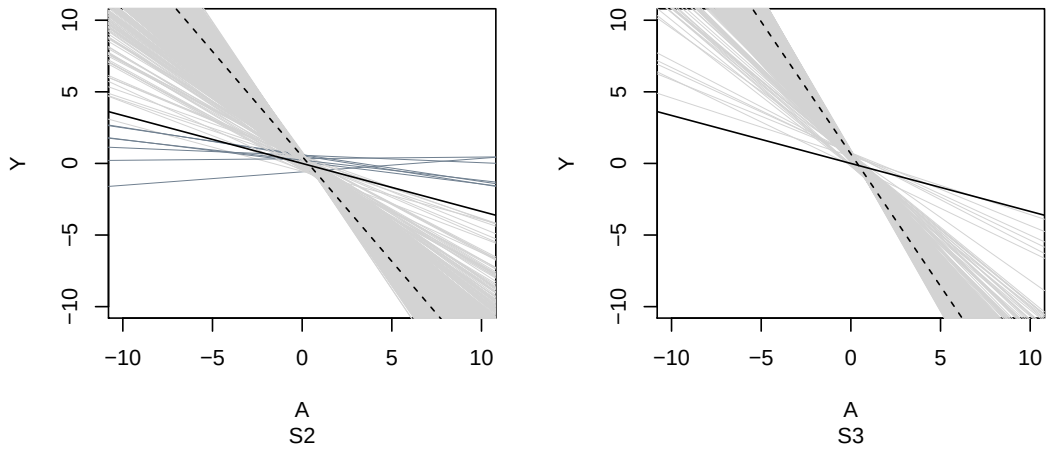




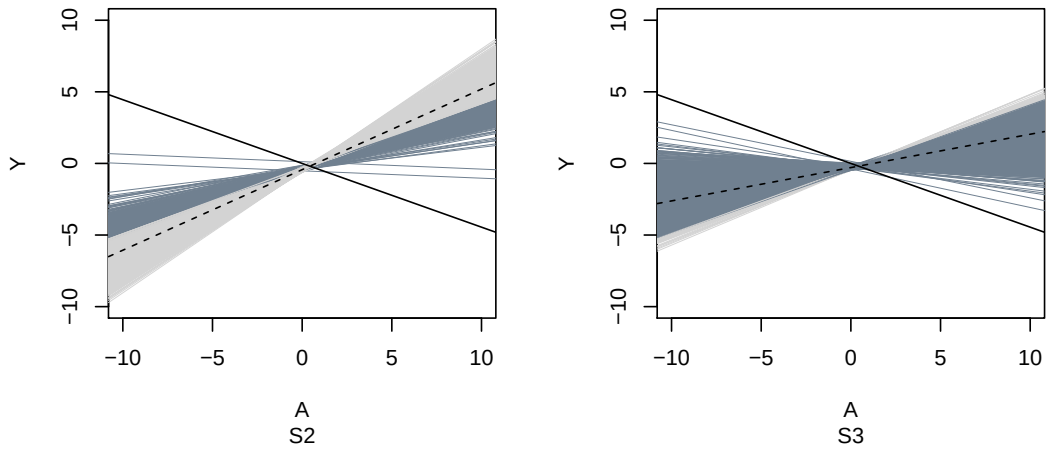
Attenuation

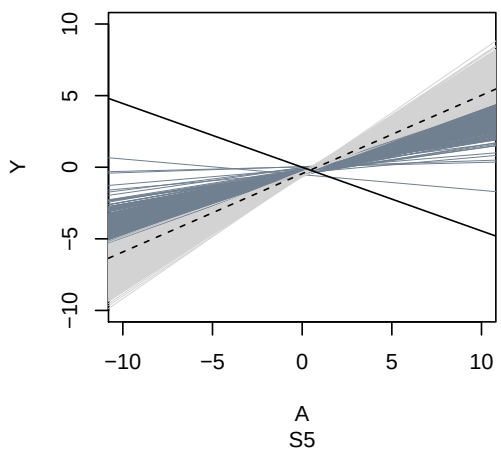


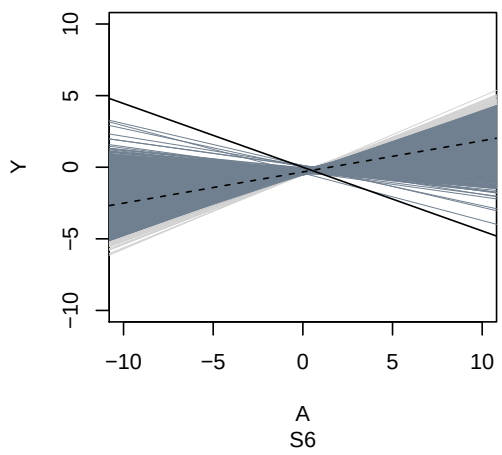
Augmentation



Sign reversal







Model of interest

- The portion of the stabilized weight which is of primary interest is the joint distribution found in the denominator of the stabilized weight:

$$p(A_i = a_i | C_i = 0, X_i = x_i) p(C_i = 0 | X_i = x_i)$$

- We consider the situation where confounders are measured with error,

$$p(A_i = a_i | C_i = 0, X_i^* = x_i^*, Z = z_i) p(C_i = 0 | X_i^* = x_i^*, Z = z_i)$$

where Z denotes observed and correctly measured confounders.

The joint distribution for the model of interest

The joint distribution of the model of interest is

$$p(A, C, X^*, X|Z = z_i; \theta)$$

where

- A denotes the treatment,
- C is binary and indicates censoring ($C = 0$, observed),
- X^* is the measured confounders,
- X is the unmeasured confounders, and
- Z are other perfectly measured covariates.
- θ is the vector of parameters.

Log-likelihood

The associated complete data log-likelihood for the model of interest is

$$\ell_c = \sum_{i=1}^{n} \log p(A_i|c=0, x_i, z_i; \theta^A) + \log p(C_i|x_i, z_i; \theta^C) + \log p(X_i^*|x_i, z_i; \theta^M) + \log p(X_i|z_i; \theta^X)$$

where

- $\theta = \{\theta^A, \theta^C, \theta^M, \theta^X\}$,
- θ^A parameterizes the treatment model,
- θ^C parameterizes the censoring mechanism,
- θ^M parameterizes the measurement error model,
- θ^X parameterizes the unobserved confounder(s) model.
- Assuming all models are uniquely parameterized.

Observed likelihood

Since X is unobserved, we should use the observed likelihood,

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \int_{\mathcal{X}} p(A_i, C_i, X_i^*, X_i | Z = z_i; \theta) d\nu_{x_i}$$
$$= \prod_{i=1}^{n} \int_{\mathcal{X}} \mathcal{L}_{ci}(\theta) d\nu_{x_i}$$

Using the observed likelihood can be more complicated than using the complete-data likelihood, so we would like to use the complete-data likelihood instead.

The approach

We use the EM algorithm, a general iterative algorithm for maximum-likelihood estimation in incomplete-data situations.

We can view measurement error as a type of missing data problem.

- The X are unobservable - missing.
- We observe X^* and we have or assume a functional relationship between X^* and X.
- We assume that all observations are mismeasured under the same measurement error model.

Basic idea of the EM algorithm

Objective

Iterative procedure to obtain maximum likelihood parameters when maximum likelihood estimation would be straightforward, but there is the additional complexity of incomplete information.

Principle

The EM algorithm is less an algorithm and more a two-step general principle.

1. E-step: Take the conditional expectation of the complete likelihood, $\ell_c(\theta|\cdot)$, given the observed data.
2. M-step: Maximize the conditional expectation with respect to the parameter.

EM Algorithm: The procedure

- Unobservable complete-data log likelihood is replaced by the conditional expectation of the complete-data log likelihood given the observed data and current parameter estimates
- For the $(t+1)^{\text{th}}$ iteration the **E-step** is

$$Q(\theta^{(t+1)}|\theta^{(t)}) = \sum_{i=1}^{n} \mathrm{E}\left[\ell_c(\theta|\cdot)|\theta^t, \text{observed}\right]$$

- For the **M-step**, choose $\theta^{(t+1)}$ such that $\theta^{(t+1)} \in \Theta$ and maximizes $Q(\theta|\theta^t)$,
  - i.e. $\theta^{(t+1)} = \text{argmax}_{\theta\in\Theta}\, Q(\theta|\theta^{(t)})$

The Generalized EM (GEM) algorithm

We can generalize the EM algorithm by modifying the M-step:

Choose $\theta \in \Theta$ such that $Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)})$

- We choose the updated parameter estimate to increase the Q-function rather than maximize it over the entire parameter space, Θ.
- This is sufficient to ensure that $\mathcal{L}(\theta^{(t+1)}) \geq \mathcal{L}(\theta^{(t)})$.
- We are not decreasing the likelihood after each GEM iteration.

The Q-function for our likelihood

We use the EM algorithm for our particular problem, thus

$$\begin{aligned} Q(\theta^{(t+1)}|\theta^{(t)}) &= \sum_{i=1}^{n} Q_i(\theta^{(t+1)}|\theta^{(t)}) \\ &= \sum_{i=1}^{n} \mathrm{E}\left[\ell_{ci}(\theta)|a_i, c_i, x_i^*, z_i; \theta^{(t)}\right] \\ &= Q(\theta^{A(t+1)}|\theta^{A(t)}) + Q(\theta^{C(t+1)}|\theta^{C(t)}) \\ &\quad + Q(\theta^{M(t+1)}|\theta^{M(t)}) + Q(\theta^{X(t+1)}|\theta^{X(t)}) \end{aligned}$$

Observations from preliminary trials

It was observed that:

- The EM algorithm would often find a ridge or plateau.
- The expectation-conditional maximization (ECM) (Meng and Rubin 1993) had similar problems.
- Evidence strongly suggested that these problems were linked to the estimation of θ^X.

Decision: Take the idea of the ECM and break the problem into smaller and simpler steps such that we gain stability in parameter estimation and we retain the desired properties of a GEM algorithm.

The basic idea of the SCEM

The sequential-conditional expectation-maximization (SCEM) algorithm is the result of breaking the problem into two simpler components.

We

1. Estimate $\{\theta^{C(t)}, \theta^{M(t)}, \theta^{X(t)}\}$ first, then
2. Estimate $\theta^{A(s)}$ while holding $\{\theta^{C(t)}, \theta^{M(t)}, \theta^{X(t)}\}$ fixed.

The stages

- The first stage, estimation of $\{\theta^C, \theta^M, \theta^X\}$, is the EM algorithm to the joint distribution

$$p(C, X^*, X|Z = z_i; \theta)$$

- The second stage, estimation of θ^A, is the EM algorithm on the conditional distribution

$$p(A|C = 0, X^* = x^*, X = x, Z = z_i, \theta^{C(t)}, \theta^{M(t)}, \theta^{X(t)}; \theta^A)$$

Q-functions for the two stages

The Q-functions for the two stages are

1. $Q(\theta^{1(t)}|\theta^{1(t-1)}) = Q(\theta^{C(t)}|\theta^{C(t-1)}) + Q(\theta^{M(t)}|\theta^{M(t-1)}) + Q(\theta^{X(t)}|\theta^{X(t-1)})$
2. $Q(\theta^{2(s)}|\theta^{2(s-1)}) = Q(\theta^{A(s)}|\theta^{A(s-1)})$

Such that $Q(\theta^{(k)}|\theta^{(k-1)}) = Q(\theta^{1(t)}|\theta^{1(t-1)}) + Q(\theta^{2(s)}|\theta^{2(s-1)})$ where $k = s + t$ and

$$Q(\theta^{A(s)}|\theta^{A(s-1)}) = \sum_{i=1}^{n} \mathrm{E}\left[\log p(A_i|c = 0, x_i, z_i; \theta^A)|a_i, c_i, x_i^*, z_i, \theta^{1(t)}; \theta^{A(s-1)}\right]$$

Immediate tasks

With this construction, there are two basic questions that need addressing,

1. Is this still a GEM, and
2. Does it search over the entire parameter space?

Showing it is a GEM

As with any EM based approach we obtain a set of initial parameter estimates, $\theta^{(0)}$. Since each component is conceived as a GEM we need only show that the sum is itself a GEM,

$$
\begin{aligned}
Q(\theta^{1(0)}|\theta^{1(0)}) + Q(\theta^{2(0)}|\theta^{2(0)}) &\leq Q(\theta^{1(1)}|\theta^{1(0)}) + Q(\theta^{2(0)}|\theta^{2(0)}) \\
&\vdots \\
&\leq Q(\theta^{1(t)}|\theta^{1(t-1)}) + Q(\theta^{2(0)}|\theta^{2(0)}) \\
&\leq Q(\theta^{1(t)}|\theta^{1(t-1)}) + Q(\theta^{2(1)}|\theta^{2(0)}) \\
&\vdots \\
&\leq Q(\theta^{1(t)}|\theta^{1(t-1)}) + Q(\theta^{2(s)}|\theta^{2(s-1)}) \\
&= Q(\theta^{(k)}|\theta^{(k-1)})
\end{aligned}
$$

Implications for having a GEM

Since we have, by definition, a GEM, we also possess the desired properties of having

- The desired convergence properties
  - Converges to a maximum, perhaps not a global one.
- Sufficient conditions for $\mathcal{L}(\theta^{(k)}) \geq \mathcal{L}(\theta^{(k-1)})$

Space-filling property

- We want to ensure that the algorithm searches over the entire parameter space.
- We will need to
  - Define unique parameterization,
  - Define functions of the parameters, and
  - Show that the convex hull of all of all feasible directions the algorithm can take at θ^k is the entire Euclidean space, $\mathbb{R}^d$, where $d = dim(\Theta)$.

Defining unique parameterization

- We say that our models are uniquely parameterized when $\theta^A, \theta^C, \theta^M, \theta^X$ are mutually orthogonal

$$\theta^i \perp \theta^j \text{ such that } i \neq j \text{ and } i, j \in \{A, C, M, X\}$$

- Thus,
  - $\theta^1 = \bigoplus_{i \in \Gamma} \theta^\Gamma$ where $\Gamma = \{A, C, M\}$.
  - $\theta = \theta^1 \oplus \theta^2$, where $\theta^2 = \theta^A$.
  - θ^1 is the orthogonal complement of θ^2.

SCEM as a "constrained" maximization

We can view the proposed SCEM algorithm as a two step procedure for which each step is maximized subject to a "constraint" which will define the subspace of Θ over which maximization occurs.

We choose to maximize $Q(\theta^{(k)}|\theta^{(k-1)})$ subject to

$$g_p(\theta) = \theta^p$$

such that $\theta^p \subseteq \theta$.

Searching in any direction from any point: Part 1

As laid out in Meng and Rubin, we want to be able to search in *any* direction, at any point in Θ, for the maximum.

- We consider the gradient at point θ^r for $g_p(\theta)$
- Under the assumption of unique parameterization as we have defined it, then $\nabla g_p(\theta)$ is full rank at $\theta^r \in \Theta$ for all r.
  - The gradient is a set of d_p elementary column vectors and a set of $d_\Theta - d_p$ zero column vectors.
  - $d_\Theta = dim(\Theta)$ and $d_p = dim(\theta^p)$

Searching in any direction from any point: Part 2

If we take $\eta \in \mathbb{R}^{d_p}$, then the column space is

$$\begin{aligned} G_p(\theta) =& \text{span}(\nabla g_p(\theta)) \\ =& \{\nabla g_p(\theta)\eta | \eta \in \mathbb{R}^{d_p}\} \end{aligned}$$

so the gradient of our constraint spans the subspace Θ^p.

This means that we can step in any direction from $\theta^r \in \Theta^p$ towards a maximum.

Searching in any direction from any point: Part 3

By construction, the intersection of $G_p(\theta)$ for all p is

$$\bigcap_p G_p(\theta^r) = \emptyset$$

for $\theta^r \in \Theta^p$ for all r.

- *The intersection of the column spaces is empty.*
- This is the empty set in $\mathbb{R}^{d_\Theta}$

Searching in any direction from any point: Part 3

Heuristically, we will take the complement in order to identify the union of the column spaces,

$$\begin{aligned}\left[\bigcap_p G_p(\theta^r)\right]^c &= \bigcup_p G_p(\theta^r)^c \\ &= \bigcup_p \{\nabla g_p(\theta)\eta | \eta \in \mathbb{R}^{d_p}\}^c \\ &= \bigcup_p \{\nabla g_p(\theta)\eta | \eta \in \mathbb{R}^{d_p}\}^\perp \\ &= \mathbb{R}^{d_\Theta}\end{aligned}$$

Implications

- By the construction of our parameter space, the union of linear hulls is itself a linear hull.
- From any point we can step in any direction in Θ in order to maximize.

We maximize over the entire Euclidean space $\mathbb{R}^{d_\Theta}$.

- i.e. we maximize over all Θ.

Technical note: Convex hull

If we define $\eta \in T_p(\theta)$ where

$$T_p(\theta) = \{\eta \in \mathbb{R}^{d_p} | \exists \{\theta_n^p\} \text{ s.t. } \eta = \lim_{n \to \infty} \frac{\theta_n^p - \theta^p}{||\theta_n^p - \theta^p||}\}$$

then we obtain a convex hull by satisfying the additional constraint that $\sum_i \eta_i = 1$ where $\eta_i = [\eta]_i$ (Meng and Rubin, 1993).

Implementation for each component

Use Monte-Carlo (MC) integration to approximate the expectation

$$\tilde{Q}(\theta^p|\theta^{p(t)}) = \sum_{i=1}^{n} \frac{1}{m_{g_i(t)}} \sum_{l=1}^{m_{g_i(t)}} \ell_c(\theta^p)$$

where $m_{g_i(t)}$ is the Monte Carlo (MC) sample size as a function of the tth step for the ith subject

Making a long story short

- Use Gibbs Adaptive Rejection Sampling (GARS) algorithm (Wild 1993)
- Implicit in this choice is a restriction to log-concave functions (Gilks 1992 deals with the exponential family)

M-step

- Maximization of $\tilde{Q}(\theta^p|\theta^{p(t)})$ is equivalent to component-wise maximization.
- In many situations, this can be done using standard software.

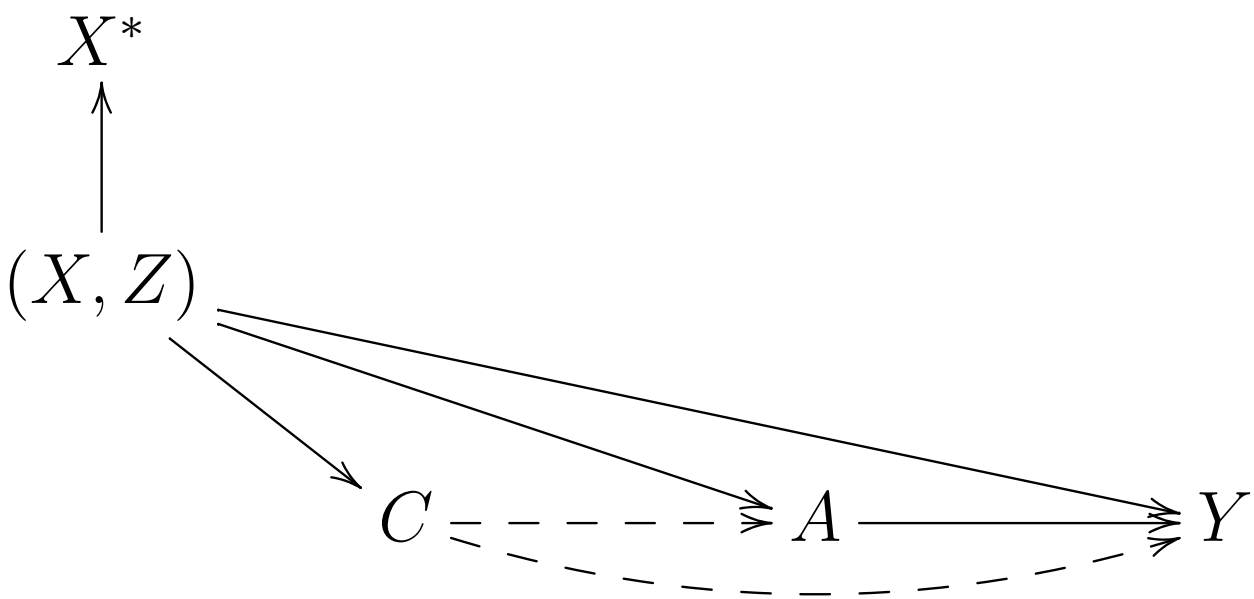
Simulation set-up: Simulation structure

- 100 Simulations.
- Sample size, n=500 for each simulation.
- Monte Carlo sample size, 2500.
- Burn-in for MC integration, 1000.
- Dissimilarity criterion for each step of the SCEM:
  1. $|\theta^{1(t)} - \theta^{1(t-1)}| \leq 0.0025$
  2. $|\theta^{2(s)} - \theta^{2(s-1)}| \leq 0.0025$

Data generation DAG

Data was generated according to the following MSM DAG.

- Dashed lines indicate that the censoring mechanism was not included as a covariate in the data generating models, but that it does affect what is observable.



Simulation set-up: Confounders

- $f(X_1, X_2) \sim MVN$.
- $\mu = (0, 0)$.
- $\sigma_{ii} = 1$ for $i = 1, 2$.
- $\sigma_{1,2} = 0.2$.

Simulation set-up: Measurement error model

- Chose X_1 to be unobservable, but has observable surrogate X_1^*
- Unbiased classic measurement error model:
  - $X_1^* = X_1 + \epsilon$
  - $\epsilon \sim N(0, \tau)$
- We are assuming τ to be known

Simulation set-up: Model parameterization

- Censoring mechanism:
  - $\text{logit}[\Pr(C = 1|X = x)] = \theta_0^C + \theta_1^C x_1 + \theta_2^C x_2$
  - where $\theta^C = (\theta_0^C, \theta_1^C, \theta_2^C) = (-3.664, 0.378, -1.881)$
- Treatment model:
  - $\text{logit}[\Pr(A = 1|X = x)] = \theta_0^A + \theta_1^A x_1 + \theta_2^A x_2$
  - where $\theta^A = (\theta_0^A, \theta_1^A, \theta_2^A) = (-0.405, 2.630, 2.307)$

Simulation execution

- Generated data as specified
- Removed X_1 from the data and censored treatments (i.e. created an observable data set)
- Assumed correction functional form for all models
- Assumed τ known
- Performed two fits:
  1. Unadjusted: Does not account for measured surrogate (naive data analysis)
  2. Adjusted: Accounts for measured surrogate (SCEM analysis)

Results for step 1: $\theta^1 = \{\theta^C, \theta^M, \theta^X\}$

Model	θ_i^C	$\hat{E}(\theta^C)(\hat{\sigma}_{\theta^C})$	$Bias(\theta^C)(SE_{\text{boot}})$	$MSE(\theta^C)$
Unadjusted	θ_0^C	-3.698 (0.362)	-0.034 (0.037)	0.133
	θ_1^C	0.289 (0.186)	-0.089 (0.018)	0.043
	θ_2^C	-1.866 (0.308)	0.015 (0.031)	0.095
Adjusted	θ_0^C	-3.718 (0.373)	-0.054 (0.038)	0.142
	θ_1^C	0.367 (0.238)	-0.011 (0.023)	0.057
	θ_2^C	-1.890 (0.319)	-0.009 (0.032)	0.102

Results for step 2: $\theta^2 = \theta^A$

Model	θ_i^A	$\hat{E}(\theta^A)(\hat{\sigma}_{\theta^A})$	$Bias(\theta^A)(SE_{\text{boot}})$	$MSE(\theta^A)$
Unadjusted	θ_0^A	-0.341 (0.153)	0.064 (0.015)	0.028
	θ_1^A	1.735 (0.174)	-0.895 (0.017)	0.832
	θ_2^A	2.027 (0.202)	-0.280 (0.019)	0.119
Adjusted	θ_0^A	-0.410 (0.202)	-0.004 (0.020)	0.041
	θ_1^A	2.704 (0.446)	0.074 (0.044)	0.204
	θ_2^A	2.361 (0.302)	0.054 (0.029)	0.094

Simulation summary

- Bias resulting from using the surrogate in a naive fashion has a bigger effect on the estimation of θ_1^A than on θ_1^C, although both are biased.
- The SCEM approach provides a nice reduction of bias for θ_1^C but has a large reduction in the magnitude of the bias for θ_1^A.
  - This is also seen in the MSE.
- Typical trade-offs seen for θ_0^C and θ_2^C with the application of the SCEM, but these trade-offs did not manifest for θ_0^A and θ_2^A.
- Simulation study suggests that the theoretical convergence properties should be similar to the EM and ECM properties.

Conclusions

- The SCEM is a variation of the ECM algorithm and and extension of the EM algorithm.
- The SCEM produces unbiased estimates of model parameters for a class of models which has constrained covariates for one of the conditional models.
  - For example, the denominator of the stabilized weight when using inverse probability of treatment weights.
- The SCEM is a GEM.
- The SCEM permits searches over the entire parameter space under the assumption that the parameters are variationally independent (i.e. unique parameterization of the models).
- The simulation results suggest that we have retained the convergence properties desired in all variations of the EM algorithm.

Next Steps

- Theoretical confirmation of the convergence properties.
- Extension of the simulation study to investigate robustness to violations of assumptions and model specifications.
- Application to MSM point-treatment context.

References

- AP Dempster, NM Laird, and DB Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm. *JRSS - B*, 39:1-38, 1977.
- WR Gilks and P Wild. Adaptive rejection sampling for Gibbs sampling. Applied Statistics, 41(2):337-348, 1992.
- P Gustafson. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. Chapman & Hall/CRC, 2004.
- MA Hernan, B Brumback, and JM Robins. Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments. *JASA*, 96(454): 440-448, 2001.
- MA Hernan and JM Robins. Estimating causal effects from epidemiological data. *BMJ*, 60(7): 578-586, 2006.
- GJ McLachlan and T Krishnan. *The EM algorithm and Extensions*. Wiley, 2008.
- XL Meng, DB Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2): 269-278, 1993.
- KM Mortimer, R Neugebauer, M van der Laan, and IB Tager. An Application of Model-Fitting Procedures for Marginal Structural Models. *AJE*, 162(4): 382-388, 2005.
- JM Robins. Association, causation, and marginal structural models. *Synthese*, 121(1):151-179, 1999.
- P Wild and WR Gilks. Algorithm as 287: Adaptive rejection sampling from log-concave density functions. Applied Statistics, 42(4):701-709, 1993.

Future work

- More extensive simulation study (higher number of replications, more scenarios for double clustering severity, programming of MCMC from scratch instead of using WinBUGS).
- Use of more informative priors for variance parameters.
- Model extension for continuous outcome: spatial component (correlation between clusters).
- Extension of current model to binary and count outcomes.
- Alternative methodologies for the analysis of continuous outcomes under double clustering: regression calibration.