Supplementary materials for: A latent variable approach to studies of gene-environment interactions in presence of multiple correlated exposures

Brisa N. Sánchez*, Shan Kang, and Bhramar Mukherjee

[1]Department of Biostatistics, University of Michigan, Ann Arbor.

**email:* brisa@umich.edu

Simulation Studies

We conducted a small scale simulation study to examine the finite sample properties of estimators under various settings of the true data generating model using $l = 1, p = 4$, and two genetic classes. Genetic class was generated as a binary variable with prevalence 0.2, similar to our data example. The exposure model parameters used for each simulation setting are given in Supplementary Materials Table 1. Since there are only two gene classes and one latent exposure, we denote the gene effect among unexposed, the exposure effect among wild types, and the interaction parameters as β_G, β_U, $\beta_{G\times U}$.

To investigate Type I error rates, data were first generated assuming $\beta_G = \beta_U = \beta_{G\times U} = 0$, and two scenarios of the G-E association: independence ($A0$) and dependence ($A3$). Similarly, to examine relative efficiency and bias, data were generated assuming G-E $A0$ or dependence $A3$, respectively, using $\beta_U = 1$, $\beta_G = \beta_{G\times U} = 2$ (equivalent to standardized effects of 0.2, 0.4, 0.4, respectively, since outcome variance was set to $\sigma^2 = 5^2$). For each simulation scenario we simulated 500 data sets of sample size 350 and obtained MLEs using software package M$plus$ (Muthén and Muthén, 2006). Parameter estimates were imported into R, where we calculated EB estimates using functions available as supplementary materials.

For each outcome model parameter we report percent bias, Bias$\% = (\beta - \overline{\widehat{\beta}})/\beta$) (or bias when $\beta = 0$), ratio of empirical variances comparing variances assuming $A1$-$A3$ to $A0$ (Var.R), empir-

ical mean squared error (MSE), and rejection probabilities (P(R)). Note that although the model is identifiable, lack of convergence was encountered because the overall ratio of sample size per parameter is low ($350/26 \approx 14$), specially for $A3$ (among those with $G = 1$, sample size per parameter is $0.20 * 350/13 \approx 5.4$). To calculate simulation results, we excluded data sets where the condition number of the information matrix was smaller than 10^{-4} or where the estimated asymptotic variance for any model parameter was greater than four times the empirical variance of the estimates. For Type I error simulations, 4.6% ($A0$) and 7.0% ($A3$) of the data sets were excluded, and 3.6% ($A0$) and 2.4% ($A3$) for the efficiency and bias simulations. Simulation studies with larger sample sizes showed no convergence problems, demonstrating that lack of convergence in the main simulation study are due to the small sample size per parameter.

References.

Muthén, L. and Muthén, B. (2006). *Mplus: Statistical Analysis with Latent Variables User's Guide*. Muthén & Muthén, Los Angeles.

Table 1: Parameters for simulations

Parameter	Type 1 Error	Efficiency	Type 1 Error	Bias
Outcome Model[a]				
β_0	1	1	1	1
β_U	0	1	0	1
β_G	0	2	0	2
$\beta_{G\times U}$	0	2	0	2
σ_ϵ	5	5	5	5
Exposure Model	A0	A0	A3	A3
Model for latent variable				
α_0	1	1	1	1
γ_g	0	0	1	1
$\Phi_{g=0}$	1	1	1	1
$\Phi_{g=1}$	1	1	1	1
Measurement Model				
Values for wildtypes				
ν_1	0.0	0.0	0.0	0.0
ν_2	1.5	1.5	1.5	1.5
ν_3	1.2	1.2	1.2	1.2
ν_4	1.0	1.0	1.0	1.0
λ_1	1.0	1.0	1.0	1.0
λ_2	0.5	0.5	0.5	0.5
λ_3	1.25	1.25	1.25	1.25
λ_4	1.25	1.25	1.25	1.25
Θ_{11}	1.86	1.86	1.86	1.86
Θ_{22}	0.52	0.52	0.52	0.52
Θ_{33}	3.25	3.25	3.25	3.25
Θ_{44}	3.65	3.65	3.65	3.65
Values for variants[b]				
ν_2			1.88	1.88
ν_3			1.50	1.50
ν_4			1.25	1.25
λ_1			1	1
λ_2			0.75	0.75
λ_3			1.88	1.88
λ_4			1.88	1.88
Θ_{11}			2.76	2.76
Θ_{22}			0.78	0.78
Θ_{33}			4.87	4.87
Θ_{44}			5.47	5.47

[a]Given residual standard deviation = 5, effect sizes are: $\beta_U = 1/5 = 0.2$ and $\beta_{G\times U} = \beta_U = 2/5 = 0.4$

[b]Values are the same as for wild types unless specified

Table 2: Main effects of lead (Model 1) and lead and iron metabolism genes (Model 2), without interaction.

	Model 1			Model 2		
Outcome Model	Est	SE	Est/SE	Est	SE	Est/SE
Intercept	3063.80	114.96	26.65	3090.45	115.18	26.83
Latent lead exposure	-54.12	25.06	-2.16	-55.03	25.14	-2.19
Maternal age	10.40	3.91	2.66	10.56	3.89	2.72
Female	-119.31	40.55	-2.94	-126.65	40.50	-3.13
Iron Genes				-96.90	49.74	-1.95
Exposure Model						
LV model						
α	15.40	0.78	19.59	15.30	0.78	19.69
Φ	11.80			11.70		
Measurement Model						
λ_2	1.03	0.05	19.50	1.03	0.05	19.71
λ_3	1.76	0.02	72.01	1.76	0.02	71.94
λ_4	2.04	0.02	88.69	2.04	0.02	88.73
Θ_{11}	1.22	0.29	4.26	1.23	0.29	4.30
Θ_{22}	0.57	0.14	4.19	0.58	0.14	4.15
Θ_{33}	0.15	0.04	3.88	0.14	0.04	3.84
Θ_{44}	0.13	0.03	3.85	0.13	0.03	3.83
Θ_{34}	0.17	0.01	11.65	0.17	0.01	11.71