Total Effects Estimation of a group of Environmental Mixtures

Xuelong Wang

August 06, 2019

A motivation example

Goal

Investigate the association between environmental exposures and a health outcome

A linear model

$$Y = \beta^T X + \epsilon$$

- Y is a heath outcome(e.g. glycohemoglobin),
- $X = (X_1, \ldots, X_p)$ are environmental exposures, e.g. PCBs,
- $\beta = (\beta_1, \ldots, \beta_p)$, $\epsilon \sim N(0, \sigma_\epsilon^2)$.

A motivation example

Variable selection and coefficient estimation (e.g. Lasso)

This approach may not work well due to :

- Low level of exposures and possibly weak effects
  - Sparsity assumption
- High correlation among the exposures
  - Collinearity

Association and Variation

Heritability

Heritability is a statistics that summarizes how much variation of a quantitative trait attribute to genetic factors

$$H^2 = \frac{\sigma_g^2}{\sigma_y^2}$$

Total effects

$$Var(\beta^T X) \text{ or } \frac{Var(\beta^T X)}{Var(Y)}$$

Note that we assume that X and/or β are random vector(s)

Model

Mixed model

$$Y = \beta_0 + \beta^T X + \epsilon \quad (1)$$

$$Y = \beta_0 + \beta^T X + X^T \Gamma X + \epsilon \quad (2)$$

Where Y is the health outcome, $X_{p\times 1}$ is the exposures, $\beta_{p\times 1}$ is the main effect, $\Gamma_{p\times p}$ are the interaction effect, ϵ is the error term. Note that X is a random vector, β and Γ could be random or fixed effects.

Goal

Specific goals

- Estimate of total main effect $Var(\beta^T X)$ and total combined effect $Var(\beta^T X + X^T \Gamma X)$
- Estimate the variance of the total effects

Estimation methods of Total effect

GCTA: Genome-wide complex trait analysis (Yang et al (2010))

- use a working mixed model to estimate the total effect
- work with $n < p$ and $n > p$ case
- covariates need to be independent to each other
- no variance estimation and inference

EigenPrism (Janson et al (2017))

- work only with $n < p$
- covariates need to be independent to each other
- provides a conservative confidence interval for estimated total effect when $n < p$

Challenge

1. Estimate the total effect of mixtures under high correlation and high dimension setup
2. Statistical inference of the estimated total effects
3. Estimate the total main and interaction effect of mixtures

Decorrelation

Total effect is invariant of linear transformation

$$Y = \beta_0 + \beta^T X + \epsilon,$$

The variance of Y explained by X is $Var(\beta^T X) = \beta^T Var(X)\beta$

$$Z = AX$$

In order to keep the same model, $\beta \rightarrow \gamma$ and $\gamma = A^{-1}\beta$

$$Y = \beta_0 + \gamma^T Z + \epsilon,$$

The variance of Y explained by Z is same as X

$$Var(\gamma^T Z) = \gamma^T Var(Z)\gamma = \beta^T Var(X)\beta = Var(\beta^T X)$$

Decorrelation

Linear transformation to remove correlation

Let $\Sigma = Var(X)$, $A = \Sigma^{-1/2}$,

$$Z = AX \Rightarrow Var(Z) = I_p,$$

Moreover,

$$Var(\beta^t X) = Var(\gamma^T Z) = \sum_{i=1}^{p} \gamma_i^2.$$

Therefore, the task is to estimate $\Sigma^{-1/2}$ correctly

How to estimate the $\Sigma^{-1/2}$

1. Spectral decomposition

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T = UDU^T$$

where X_i with $p \times 1$ and U is $p \times p$ matrix, D is a diagonal matrix with $diag(D) = (d_1, \ldots, d_p)$,

$$\hat{\Sigma}^{-1/2} = UD^{-1/2}U^T$$

where $diag(D^{-1/2}) = (d_1^{-1/2}, \ldots, d_p^{-1/2})$

Simulation setup

Data generating model

$$Y = X\beta + \epsilon$$

- Y is a $n \times 1$ vector, X is a $n \times p$ matrix, β is a $p \times 1$ vector
- $n = 200, 500, 1000$, $p = 500$
- X follows Normal or χ^2_1 and $Var(X) = \Sigma_{un}$
- $\epsilon \sim N(0, \sigma^2_\epsilon)$
- $\frac{Var(X\beta)}{Var(Y)} = 0.5$, $Var(X\beta) = 10$

Simulation result

Table 1: Simulation result without decorrelation

x_dist	n	effect_EP	effect_GCTA	var_EP	var_GCTA	MSE_EP	MSE_GCTA
chi	200	13.89	13.58	15.2	14.25	30.12	26.89
	500	14.66	14.62	7.11	7.36	28.66	28.6
	1000	NA	14.75	NA	2.37	NA	24.93
normal	200	13.48	13.77	21.55	21.06	33.24	35.05
	500	14.57	14.46	6.88	6.32	27.68	26.11
	1000	NA	15.19	NA	2.79	NA	29.7

Table 2: Simulation result with decorrelation by spectral decomposition

x_dist	n	effect_EP	effect_GCTA	var_EP	var_GCTA	MSE_EP	MSE_GCTA
chi	200	49.87	22.9	27.96	634.85	1615.79	794.93
	500	20.12	12.82	2.78	95.68	105.13	102.62
	1000	NA	**9.8**	NA	**0.97**	NA	**1**
normal	200	51.54	24.2	19.79	677.04	1744.29	872.02
	500	20.05	10.2	1.77	101.69	102.75	100.69
	1000	NA	**9.95**	NA	**0.96**	NA	**0.95**

Sigularity of $\hat{\Sigma}$ when $n < p$

If $n < p$, $\hat{\Sigma}$ is a unbiased and consistent estimator of Σ. However, $\hat{\Sigma}$ will be singular when $n < p$, which means some of $d_i's$ are zeros. So that is the reason $\hat{\Sigma}^{-1/2}$ are not stable.

Solutions

1. Use historical data to get a good estimation of $\hat{\Sigma}$
2. Use methods for large (High dimension) covariance matrix and precision matrix estimation
   - Sparse precision matrix estimation method
   - Factor model

Simulation result: Use historical data and spectral decompostion

$$\hat{\Sigma}_h = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (X_i - \bar{X})(X_i - \bar{X})^T$$

Where n_h is the sample size of historical data.

Table 3: Simulation result with decorrelation by spectral decomposition

x_dist	n	effect_EP	effect_GCTA	var_EP	var_GCTA	MSE_EP	MSE_GCTA
chi	200	**9.99**	**9.67**	**6.04**	**6.43**	**5.98**	**6.47**
	500	**9.83**	**9.83**	**2.93**	**2.85**	**2.93**	**2.85**
	1000	NA	**9.72**	NA	**0.75**	NA	**0.82**
normal	200	**9.48**	**9.13**	**9.37**	**11.02**	**9.55**	**11.67**
	500	**10.05**	**9.88**	**2.19**	**2.19**	**2.17**	**2.18**
	1000	NA	**10.05**	NA	**0.89**	NA	**0.89**

Large covariance matrix estimation

Sparse precision matrix estimation:Glasso

$$\hat{\Theta} = \underset{\Theta=(\theta_{ij})_{p\times p}}{\arg\min} \left\{ tr(\hat{\Sigma}\Theta) + \log|\Theta| + \sum_{i\neq j} P(|\theta_{ij}|) \right\}$$

where $\Theta = \Sigma^{-1}$ and assume it is sparse P is the penalty.

Factor model

$$X_t = Bf_t + \mu_t$$

Where $B = (b_1, \ldots, b_p)^T$, f_t with $p \times 1$ is the common factor and $\mu_t = (\mu_{1t}, \ldots, \mu_{pt})^T$.

$$Var(X_t) = Bcov(f_t)B^T + \Sigma_\mu$$

Where $\Sigma_\mu = Var(\mu_t)$ is sparse.

Restricted to PCBs covariance structures

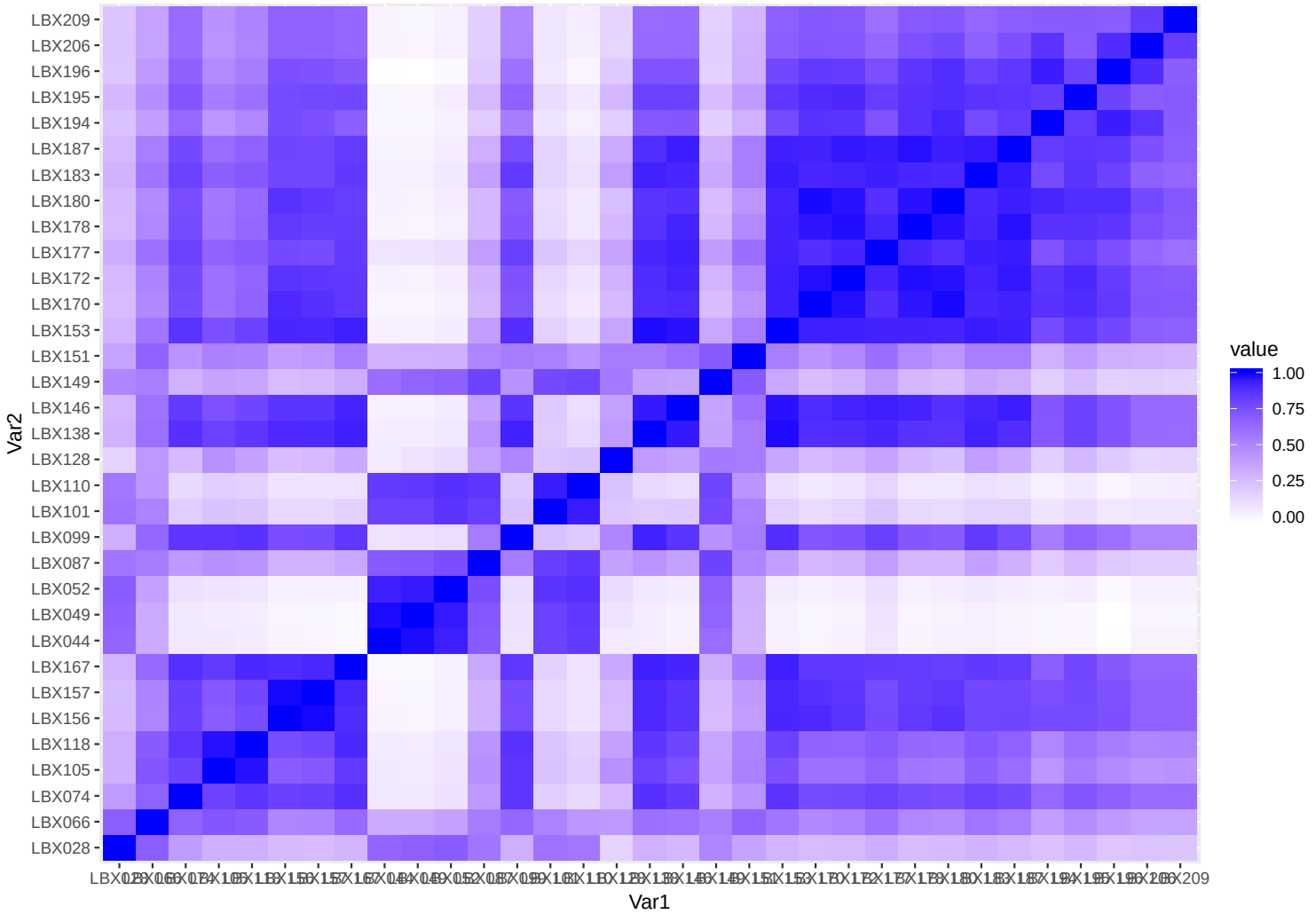

Figure 1: Sample covariance correlation of PCBs from 1999 - 2013

Restricted to PCBs covariance structures

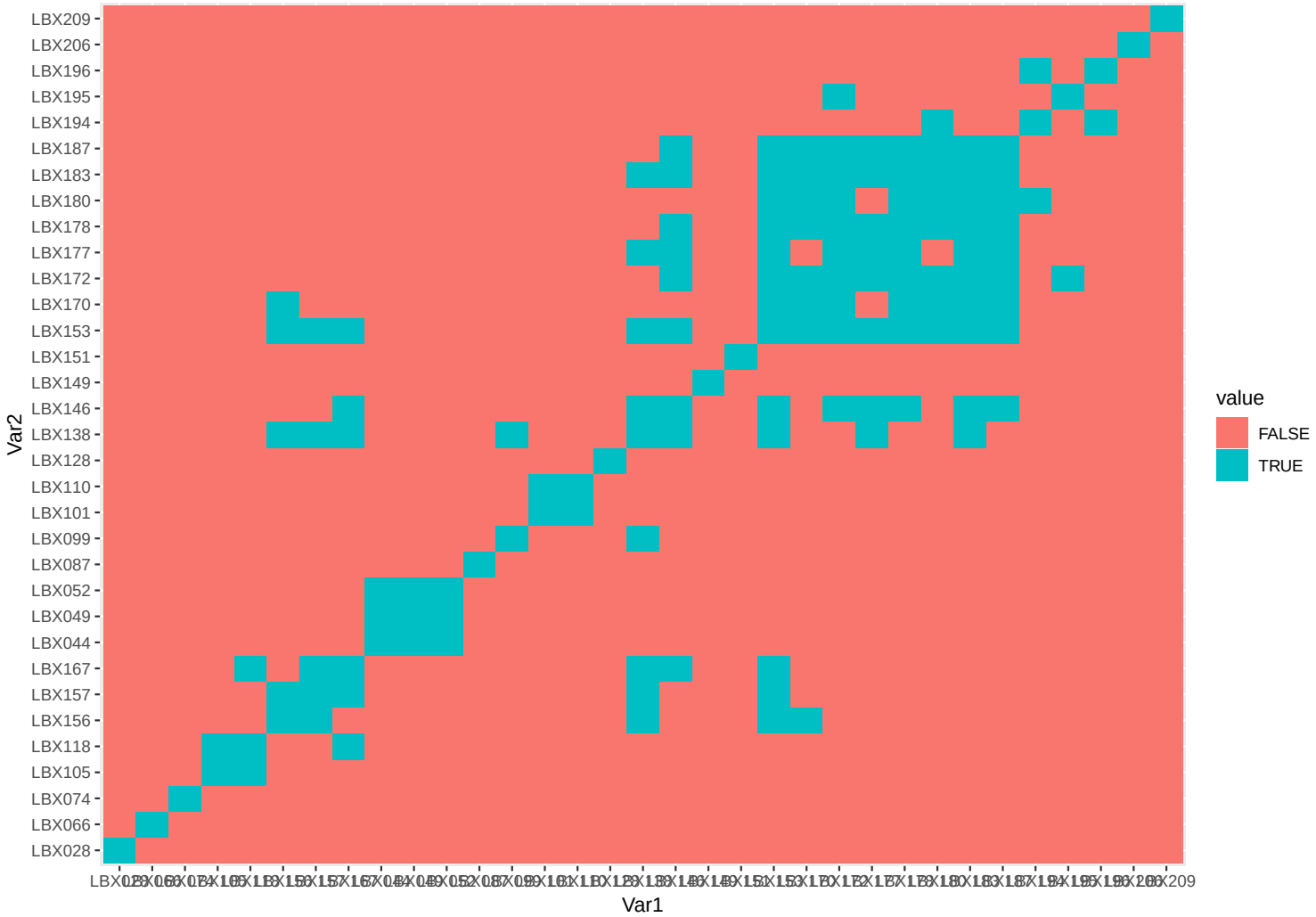

Figure 2: all the correlation > 0.9

Variance estimation

Jackknife variance estimation

$$\hat{Var}(\beta^T X) = \hat{\theta}$$

1. Sub-sample X_d a $(n-d) \times p$ matrix and Y_d is a $(n-d) \times 1$ from X and Y
2. Use the X_d, Y_d to fit the model and get the estimation $\hat{\theta}_{-d}$
3. Iterate the whole process S times and get the sub-sampling-variance as $Var(\hat{\theta}) = \frac{n-d}{d} \frac{1}{S} \sum_s (\hat{\theta}_{-d_s} - \hat{\theta}_.)^2$, where $\hat{\theta}_. = \frac{\sum_s \hat{\theta}_{-d_s}}{S}$

Note if we set d $= 1$, then we will have the leave-1-out estimator.

Simulation result

Leave-1 out method may be the best choice.

- $n = 500,\ p = 1000$
- $X \sim N(0, I)$
- Nominal coverage rate is 80%

Table 4

x_dist	method	delete	effect	var	var_jack	CI_width_sub	coverage_sub	CI_width	coverage	var_diff_ratio
normal	EigenPrism	0.7	10.14	4.24	12.67	9.09	0.96	5.71	0.85	1.99
		0.4			8.76	7.56	0.93			1.07
		0.1			7.55	7.01	0.91			0.78
		1			**7.48**	**6.98**	**0.9**			**0.76**
	GCTA	0.7	10.05	4.37	12.78	9.13	0.97	NA	NA	1.92
		0.4			9.34	7.79	0.94	NA	NA	1.14
		0.1			7.71	7.06	0.91	NA	NA	0.76
		1			**7.52**	**6.98**	**0.91**	NA	NA	**0.72**

Simulation result

The bias will be reduced when n is increasing.

- $n = 500, 1000, 2000,\ p = 500$
- X follows independent normal or chi
- Nominal coverage rate is 80%

Table 5: Simulation result of GCTA with sub-sampling

x_dist	n	var	var_jack	CI_width_sub	coverage_sub	var_diff_ratio
normal	500	2.5	5.55	6.01	0.92	1.22
	1000	0.9	1.5	3.13	0.89	0.67
	2000	0.31	0.46	1.73	0.85	0.48
chi	500	2.45	5.51	5.99	0.92	1.25
	1000	0.81	1.5	3.13	0.89	0.85
	2000	0.28	0.47	1.75	0.89	0.68

Summary and future work

Summary

1. Unbiased total effect by appropriate linear transformation
2. conservative variance estimation of total effect by Jackknife method

Future work

1. Precision matrix estimation under high dimension setup
2. Variance estimation bias correction