Bayesian Model Averaging Estimator for Binary Outcome, Treatment and Confounders

Yunlong Nie

Department of Stat, UBC

July 31, 2013

Outline

Background

Double Robustness

Bayesian Model Averaging

Simulation & Conclusion

Potential Outcome Framework

Simplest situation: observational study, $d_{1:n}$

Potential Outcome Framework

Simplest situation: observational study, $d_{1:n}$

- Outcome Y

Potential Outcome Framework

Simplest situation: observational study, $d_{1:n}$

- Outcome Y
- Treatment X : 0 unexposed; 1 exposed

Potential Outcome Framework

Simplest situation: observational study, $d_{1:n}$

- Outcome Y
- Treatment X
- Confounders $\mathbf{C} = (C_1, \ldots, C_r)$, support $\mathcal{C}$, $|\mathcal{C}| = 2^r$

Potential outcomes:

- Y_1
- Y_0

Potential outcomes:

- Y_1 what would happen if a unit is treated
- Y_0

Potential outcomes:

- Y_1
- Y_0

$$Y = (1 - X)Y_0 + XY_1$$

Objective: the *causal* effect of treatment X

Objective: the *causal* effect of treatment X

- mean if **entire** pop exposed - mean if entire pop unexposed

$$\Delta \equiv \mu_1 - \mu_0 = E(Y_1) - E(Y_0)$$

Challenge

Challenges:

- Y_1 or Y_0 is missing

Challenge

Challenges:

- Y_1 or Y_0 is missing

$$Y = (1 - X)Y_0 + XY_1$$

leads to $E(Y|X = 1) = E(Y_1|X = 1) \neq E(Y_1)$

Solution

Assumption (i):

$$(Y_0, Y_1) \perp\!\!\!\perp X$$

Solution

Assumption (i):

$$(Y_0, Y_1) \perp\!\!\!\perp X$$

exchangeability

Solution

Assumption (i):

$$(Y_0, Y_1) \perp\!\!\!\perp X$$

Too strong?

Solution for observational study

Assumption (ii):

$$(Y_0, Y_1) \perp\!\!\!\perp X | \mathbf{C}$$

Solution for observational study

Assumption (ii):

$$(Y_0, Y_1) \perp\!\!\!\perp X|\mathbf{C}$$

leads to $E(Y_0) = E\Big\{E(Y|X=0,\mathbf{C})\Big\}$

Solution for observational study

Assumption (ii):

$$(Y_0, Y_1) \perp\!\!\!\perp X | \mathbf{C}$$

conditional exchangeability

Solution for observational study

Assumption (ii): “no-unmeasured-confounders”

$$(Y_0, Y_1) \perp\!\!\!\perp X | \mathbf{C}$$

Unverifiable

Techiques

- Regression
- Inverse-probability-weighting (IPW)
- Double Robustness

Regression

Outcome model:

$$m_x(\mathbf{C}; \beta_x) = E(Y|X = x, \mathbf{C}), x \in \{0, 1\}$$

e.g., 4 confounders. β_x will have 6 components including intercept.

$$\hat{\mu}_{x,reg} = \frac{1}{N} \sum_{i=1}^{N} m_1(\mathbf{c_x}, \hat{\beta}_x).$$

$\hat{\mu}_{0,reg}$ and $\hat{\mu}_{1,reg}$ consistent to μ_0 and μ_1

IPW

Treatment model:

$$\pi(C,\alpha) = Pr(X=1|C)$$

e.g., 4 confounders. α will have 5 components including intercept.

$$\hat{\mu}_{1,ipw} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i y_i}{\hat{\pi}(\mathbf{c}_i,\hat{\alpha})},$$

$$\hat{\mu}_{0,ipw} = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-x_i)y_i}{1-\hat{\pi}(\mathbf{c}_i,\hat{\alpha})}.$$

Double Robustness

$$\widehat{\Delta}_D = \hat{\mu}_{1,D} - \hat{\mu}_{0,D}$$

$$\hat{\mu}_{1,D} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i x_i - (x_i - \hat{\pi}(c_i))\hat{m}_1(c_i)}{\hat{\pi}(c_i)}$$

$$\hat{\mu}_{0,D} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i(1 - x_i) + (x_i - \hat{\pi}(c_i))\hat{m}_0(c_i)}{(1 - \hat{\pi}(c_i))}$$

What "Double" means?

$$\hat{\mu}_{1,D} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i x_i - (x_i - \hat{\pi}(c_i))\hat{m}_1(c_i)}{\hat{\pi}(c_i)}$$

What "Double" means?

$$\hat{\mu}_{1,D} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i x_i - (x_i - \hat{\pi}(c_i))\hat{m}_1(c_i)}{\hat{\pi}(c_i)}$$

- if $m_1(C, \beta)$ is correctly specified,

$$\begin{aligned}\hat{\mu}_{1,D} &= \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{m}_1(c_i))x_i + \hat{\pi}(c_i)\hat{m}_1(c_i)}{\hat{\pi}(c_i)} \\ &\approx \frac{1}{n}\sum_{i=1}^{n} \hat{m}_1(c_i) \equiv \hat{\mu}_{1,reg}\end{aligned}$$

What "Double" means?

$$\hat{\mu}_{1,D} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i x_i - (x_i - \hat{\pi}(c_i))\hat{m}_1(c_i)}{\hat{\pi}(c_i)}$$

- if $\pi(C, \alpha)$ is correctly specified,

$$\hat{\mu}_{1,D} \approx \frac{1}{n}\sum_{i=1}^{n} \frac{y_i x_i}{\hat{\pi}(c_i)} \equiv \hat{\mu}_{1,ipw}$$

Another View

$$\Delta = \sum_{c \in \mathcal{C}} Pr(C = c)\Big(Pr(Y = 1|X = 1, c) - Pr(Y = 1|X = 0, c)\Big)$$
$$= \sum_{c \in \mathcal{C}} q_c(p_{c1} - p_{c0})$$

- $q_c = Pr(C = c)$
- $p_{cx} = Pr(Y = 1|X = x, c = c)$

Another View

$d_{1:n} : (Y, X, \mathbf{C}) \to 2 \times 2 \times |\mathcal{C}|$ table with cell counts as

$$n_{cxy} = \sum_{i=1}^{n} I(c_i = c, x_i = x, y_i = y), c \in \mathcal{C}$$

$\hat{P}r()$: sample proportion to the table.
e.g.

$$\hat{P}r(Y = 1 | X = x, C = c) = \frac{n_{cx1}}{n_{cx1} + n_{cx0}} = \frac{n_{cx1}}{n_{cx\bullet}}$$

$$\hat{P}r(C = c) = \frac{n_{c\bullet\bullet}}{n} = \frac{n_{c1\bullet} + n_{c0\bullet}}{n}$$

Another View

$$\hat{\mu}_{1,D} = \sum_{c \in \mathcal{C}} \hat{P}r(C = c) \Big\{ w_1(c)\hat{m}_1(c) + (1 - w_1(c))\hat{P}r(Y = 1 | X = 1, C = c) \Big\},$$

where

$$w_1(c) = 1 - \frac{\hat{P}r(X = 1 | C = c)}{\hat{\pi}(c)}.$$

Another View

$$\hat{\mu}_{1,D} = \sum_{c \in \mathcal{C}} \hat{Pr}(C = c) \Big\{ w_1(c)\hat{m}_1(c) + (1 - w_1(c))\hat{Pr}(Y = 1 | X = 1, C = c) \Big\},$$

where

$$w_1(c) = 1 - \frac{\hat{Pr}(X = 1 | C = c)}{\hat{\pi}(c)}.$$

- **compromises** between parametric and non-parametric <u>outcome</u> models

Another View

$$\hat{\mu}_{1,D} = \sum_{c \in \mathcal{C}} \hat{Pr}(C = c) \Big\{ w_1(c) \hat{m}_1(c) + (1 - w_1(c)) \hat{Pr}(Y = 1 | X = 1, C = c) \Big\},$$

where

$$w_1(c) = 1 - \frac{\hat{Pr}(X = 1 | C = c)}{\hat{\pi}(c)}.$$

- **compromises** between parametric and non-parametric outcome models
- treatment model controls the weighting

New Scheme for Compromise

$$\hat{\Delta}_B = w_B \hat{\Delta}_{reg} + (1 - w_B) \hat{\Delta}_S$$

- w_B posterior probability that parametric model is correct.

New Scheme for Compromise

$$\hat{\Delta}_B = w_B \hat{\Delta}_{reg} + (1 - w_B)\hat{\Delta}_S$$

- w_B posterior probability that parametric model is correct.
-

$$\begin{aligned}\hat{\Delta}_{reg} &= \frac{1}{n}\sum_{i=1}^{n}\left\{m_1(c_i;\hat{\beta}) - m_0(c_i;\hat{\beta})\right\} \\ &\equiv \sum_{c\in\mathcal{C}} \hat{Pr}(C = c)\left\{m_1(c;\hat{\beta}) - m_0(c;\hat{\beta})\right\}\end{aligned}$$

Non-parametric Part: Saturated

$$\hat{\Delta}_S = \sum_{c \in \mathcal{C}} \hat{Pr}(C = c)\left(\hat{Pr}(Y = 1|X = 1, c) - \hat{Pr}(Y = 1|X = 0, c)\right)$$

$$\Delta = \sum_{c \in \mathcal{C}} q_c(p_{c1} - p_{c0})$$

- $q_c = Pr(C = c)$ with the prior $\mathbf{q} \sim Dirichlet(k_q, \cdots, k_q)$.
- $p_{cx} = Pr(Y = 1|X = x, c = c)$ with prior $p_{cx} \sim Beta(k_p, k_p)$
- q and each p_{cx} *posteriori* independent

Non-parametric Part: Saturated

$$
\begin{aligned}
\hat{\Delta}_{S0} &= \sum_{c \in \mathcal{C}} E(q_c|d_{1:n})(E(p_{c1}|d_{1:n}) - E(p_{c0}|d_{1:n})) \\
&= \sum_{c \in \mathcal{C}} \frac{k_q + n_{c\bullet\bullet}}{|\mathcal{C}|k_q + n} \left[\frac{k_p + n_{c11}}{2k_p + n_{c1\bullet}} - \frac{k_p + n_{c01}}{2k_p + n_{0\bullet}} \right]
\end{aligned}
$$

27

Bayesian Model Averaging

$$w_B = \frac{\kappa f_{reg}(y|x,c)}{\kappa f_{reg}(y|x,c) + (1-\kappa) f_{S0}(y|x,c)},$$

where κ prior probability of correctness of parametric model

$$\hat{\Delta}_B = w_B \hat{\Delta}_{reg} + (1 - w_B)\hat{\Delta}_{S0}$$

Simulation 1: data-generating

- $\mathbf{C} = (C_1, C_2, C_3, C_4)$:
  thresholding correlated standard normals (correlation equals $\rho = 0.3$) at zero

Simulation 1: data-generating

- $\mathbf{C} = (C_1, C_2, C_3, C_4)$:
- X

$$\text{logit}Pr(X = 1|\mathbf{C}) = \alpha_0 + \alpha_1(C_1 + C_2 + C_3 - 1.5),$$

$$\alpha_1 = \frac{\text{logit}(0.5) - \text{logit}(0.05)}{4},$$
$$\alpha_0 = \frac{\text{logit}(0.5) + \text{logit}(0.05)}{2},$$

Simulation 1: data-generating

- $\mathbf{C} = (C_1, C_2, C_3, C_4)$:
- X

$$\text{logit}Pr(X = 1|\mathbf{C}) = \alpha_0 + \alpha_1(C_1 + C_2 + C_3 - 1.5),$$

- Y

$$\text{logit}Pr(Y = 1|X, \mathbf{C}) = \gamma_0 + \gamma_1 X + \lambda X(C_1 C_2 - \mu_{12})$$

$\gamma_0 = -1.2, \gamma_1 = 0.7$

Simulation 1: data-generating

- $\mathbf{C} = (C_1, C_2, C_3, C_4)$:
- X

$$\text{logit}Pr(X = 1|\mathbf{C}) = \alpha_0 + \alpha_1(C_1 + C_2 + C_3 - 1.5),$$

- Y

$$\text{logit}Pr(Y = 1|X, \mathbf{C}) = \gamma_0 + \gamma_1 X + \lambda X(C_1 C_2 - \mu_{12})$$

- $n = 100, 200, 500$

34

Simulation 1: estmators

- $\widehat{\Delta}_{reg}$: fitted with only main effect: $X, C_1, \ldots, C_4$
- $\widehat{\Delta}_{D}$: fitted with only main effect
- $\widehat{\Delta}_{S0} : k_p = k_q = 1$
- $\widehat{\Delta}_{B} : \kappa = 0.5$

Simulation 1: estmators

- $\widehat{\Delta}_{reg}$: fitted with only main effect: $X, C_1, \ldots, C_4$
- $\widehat{\Delta}_D$: fitted with only main effect
- $\widehat{\Delta}_{S0} : k_p = k_q = 1$
- $\widehat{\Delta}_B : \kappa = 0.5$
- True value

$$\Delta = E\{E(Y|X=1,C) - E(Y|X=0,C)\}$$

Note that

$$Pr(C=c) = E\left(\Phi\{(\rho^{-1}-1)^{-1/2}Z\}^{r-g(c)}[1-\Phi\{(\rho^{-1}-1)^{-1/2}Z\}]^{g(c)}\right),$$

where $g(c) = \sum_{j=1}^{r} C_j, Z \sim N(0,1)$.

Simulation 1

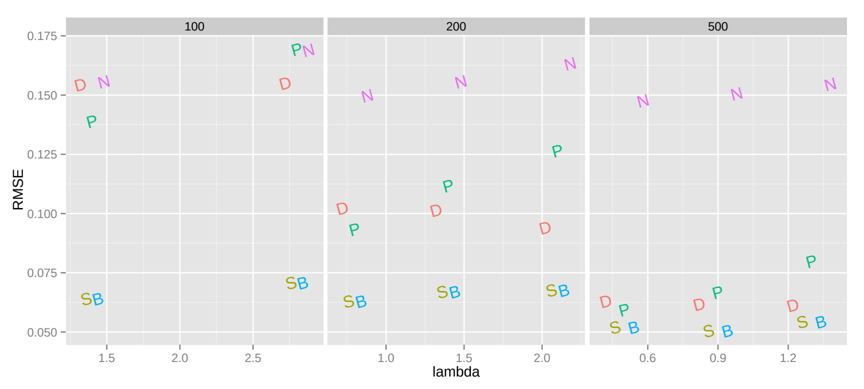


Figure : Root-mean-squared error of estimates $D = \widehat{\Delta}_D$, $S = \widehat{\Delta}_{S0}$, $P = \widehat{\Delta}_{reg}$, $B = \widehat{\Delta}_B$. N marks the true value of Δ.

Simulation 1

$$\text{logit}Pr(X = 1|\mathbf{C}) = \alpha_0 + \alpha_1 C_1 C_2 C_3$$
$$\alpha_0 = \text{logit}(0.05), \alpha_1 = \text{logit}(0.5) - \alpha_0$$

Simulation 1

$$\text{logit}Pr(X = 1|\mathbf{C}) = \alpha_0 + \alpha_1 C_1 C_2 C_3$$
$$\alpha_0 = \text{logit}(0.05), \alpha_1 = \text{logit}(0.5) - \alpha_0$$

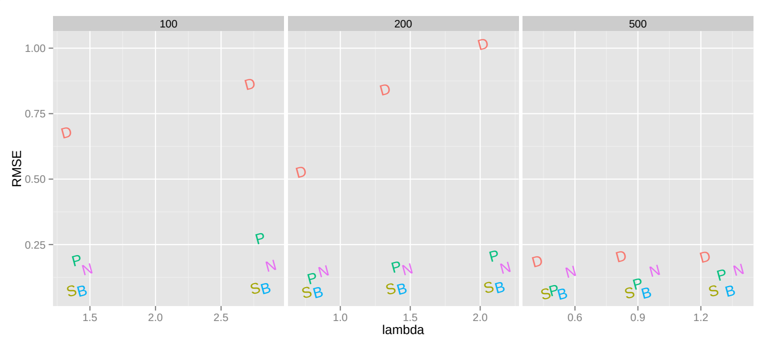


Figure : Root-mean-squared error of estimates $D = \widehat{\Delta}_D, S = \widehat{\Delta}_S, P = \widehat{\Delta}_{reg}, B = \widehat{\Delta}_B$. N marks the true value of Δ.

Limitation: sparisty

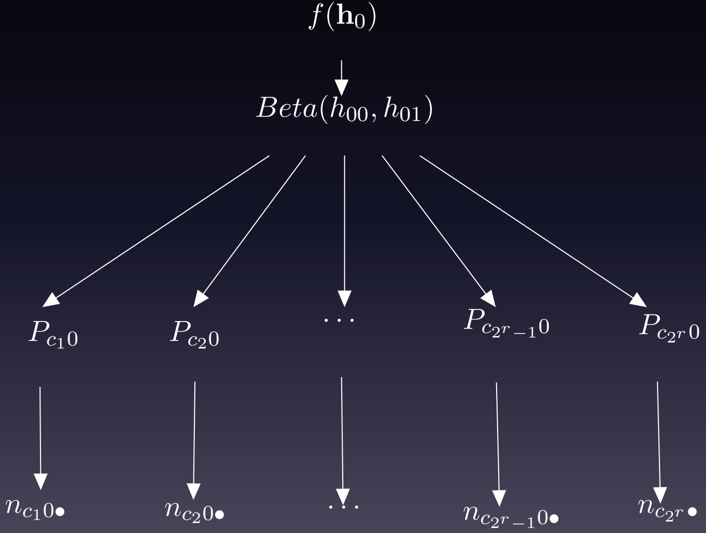
Non-sparsity:

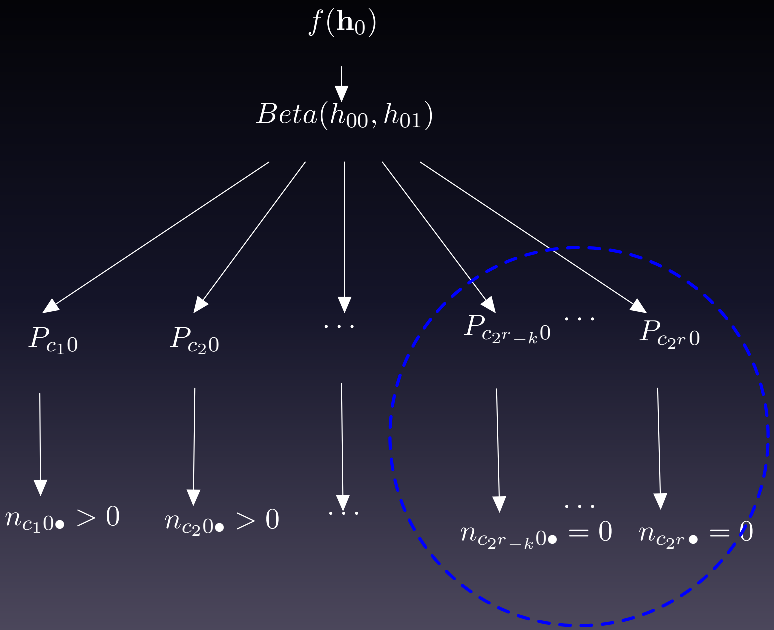
$$\hat{Pr}(C = c) > 0 \Rightarrow 1 > \hat{Pr}(X = 1|C = c) > 0$$

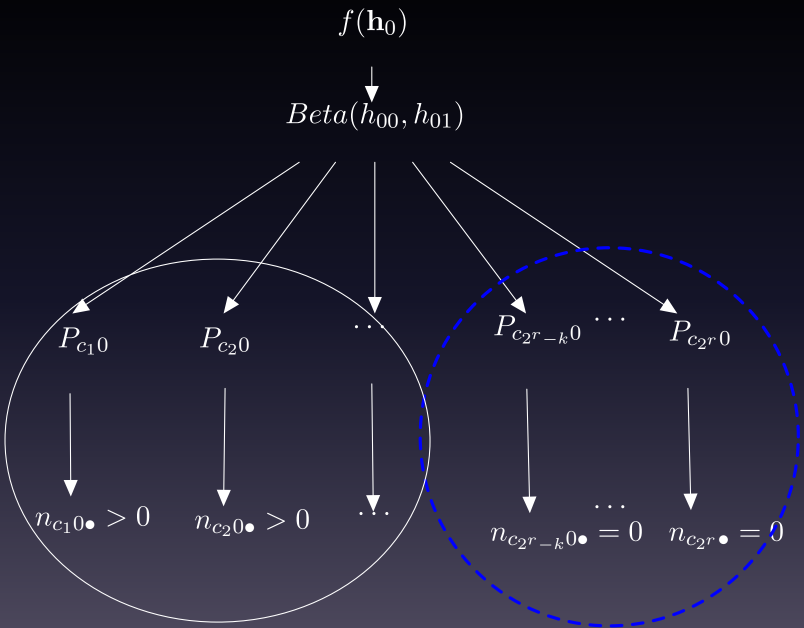
$$n_{c\bullet\bullet} > 0 \Rightarrow n_{c1\bullet} > 0, n_{c0\bullet} > 0$$

Sparsity : 21
some $n_{c1\bullet}$ or $n_{c0\bullet}$ zero, inflating the posterior variance of $\hat{\Delta}_{S0}$

Hierarchical Modelling




$f(\mathbf{h}_0)$
$Beta(h_{00}, h_{01})$
$P_{c_1 0}$
$P_{c_2 0}$
$\cdots$
$P_{c_{2^r-k} 0}$
$\cdots$
$P_{c_{2^r} 0}$
$n_{c_1 0 \bullet} > 0$
$n_{c_2 0 \bullet} > 0$
$\cdots$
$n_{c_{2^r-k} 0 \bullet} = 0$
$\cdots$
$n_{c_{2^r} \bullet} = 0$


$f(\mathbf{h}_0)$
$Beta(h_{00}, h_{01})$
$P_{c_1 0}$
$P_{c_2 0}$
$\cdots$
$P_{c_{2^r-k} 0}$
$\cdots$
$P_{c_{2^r} 0}$
$n_{c_1 0 \bullet} > 0$
$n_{c_2 0 \bullet} > 0$
$\cdots$
$\cdots$
$n_{c_{2^r-k} 0 \bullet} = 0$
$n_{c_{2^r} \bullet} = 0$

Hierarchical Modelling

- Stage1:

$$P_{cx} \equiv Pr(Y=1|X=x, C=c) \sim Beta(h_{x0}, h_{x1})$$

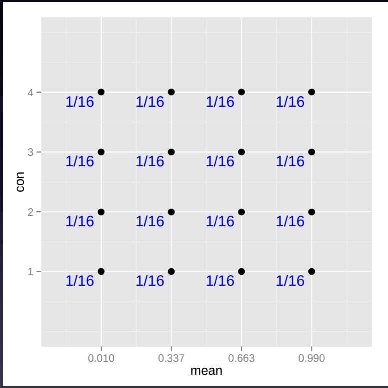
hyperparameter $\mathbf{h}_x = (h_{x0}, h_{x1})$ when $X = x, x \in \{0, 1\}$

- Stage2:

$$\mathbf{h}_x \sim f_x$$

HyperPrior

- f_x a discrete Grid G_x
- "mean":
  $m_x \equiv h_{x0}/h_{x0} + h_{x1}$
- "concentration":
  $c_x \equiv h_{x0} + h_{x1}$;
  $m_x \rightarrow$, $c_x \uparrow$, var $\uparrow$



Hierarchical Saturated Estimator

$$\hat{\Delta}_{Sh} = \sum_{c \in \mathcal{C}} E(q_c|d_{1:n})(E(p_{c1}|d_{1:n}) - E(p_{c0}|d_{1:n}))$$

$$E(q_c|d_{1:n}) = \frac{k_q + n_{c\bullet\bullet}}{|C|k_q + n}$$

$$E(p_{c1}|d_{1:n}) = \sum_{\mathbf{h}_1 \in G_1} \frac{h_{10} + n_{c11}}{h_{10} + h_{11} + n_{c1\bullet}} f(\mathbf{h}_1|d_{1:n})$$

$$E(p_{c0}|d_{1:n}) = \sum_{\mathbf{h}_0 \in G_0} \frac{h_{00} + n_{c01}}{h_{00} + h_{01} + n_{c0\bullet}} f(\mathbf{h}_0|d_{1:n})$$

Simulation 2

- Data-generating scheme: Same as Simulation 1[23]
- Sample size fixed : $n = 500$.
- Number of confounders, r : 4 to 18. $> 9 \rightarrow$ sparsity: $2^9 = 512 > 500$ different C strata

Simulation 2

- $\widehat{\Delta}_{reg}$: fitted with only main effect: $X, C_1, \ldots, C_r$
- $\widehat{\Delta}_{S0}$: $k_p = k_q = 1$
- $\widehat{\Delta}_{S1}$:
    - $G_0 \& G_1$: mean (0.01,0.99) length of 20; con (1,20) length of 20. In total, 22×22 points

Simulation2 res

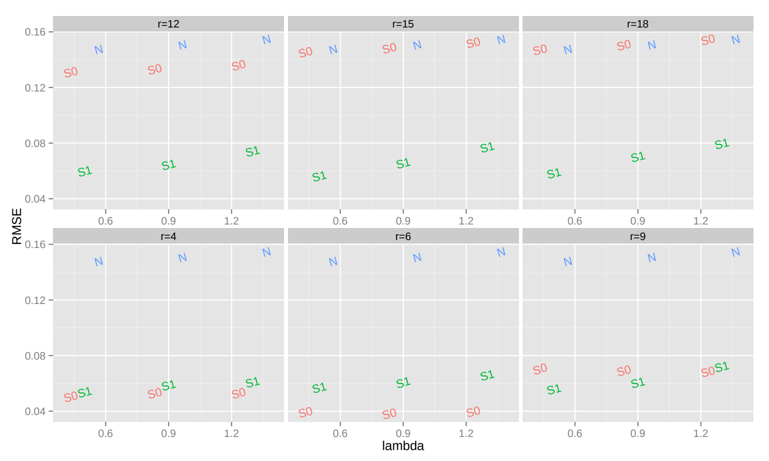
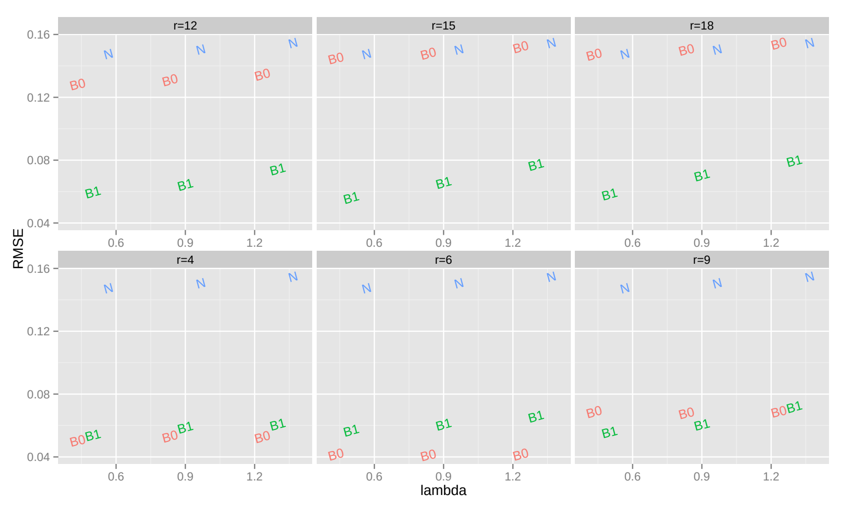


Figure : Root-mean-squared error of estimates $\widehat{\Delta}_{S0}, \widehat{\Delta}_{S1}$. 'N' marks the true value of Δ. Each panel is simulated under a different number of confounders ranging from 4 to 18.

Simulation2 res



- DR estimator : parametric and non-parametric outcome models

- DR estimator : parametric and non-parametric outcome models
- Bayesian model averaging estimator: a “cousin” of DR estimator

- DR estimator : parametric and non-parametric outcome models
- Bayesian model averaging estimator: a “cousin” of DR estimator
- Hierarchical structure helps when dealing with sparse data

- DR estimator : parametric and non-parametric outcome models
- Bayesian model averaging estimator: a “cousin” of DR estimator
- Hierarchical structure helps when dealing with sparse data

Future Work

- Continuous outcome, treatment and confounders
- Real data analysis
- R packages: https://github.com/YunlongNie/BayDR

Future Work

- Continuous outcome, treatment and confounders
- Real data analysis
- R packages: https://github.com/YunlongNie/BayDR

Thanks!