An Exhuastive Search Based Patient Subgroup Identification Approach

Lin Qiu, PSU
Jie Cheng(Mentor), AbbVie

Statistical Research Intern
Data and Statistical Science, AbbVie Inc.

December 5, 2019

Background

Figure: Newgard (2017)

AI can discover new drugs because biology is messy and complex, not in spite of that fact, says Andrew Hopkins, AbbVie (spectrum.iee.org,2018)

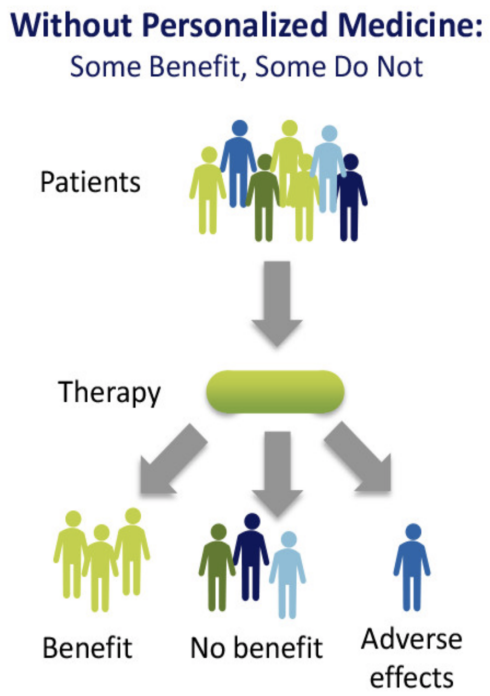
Disclosure

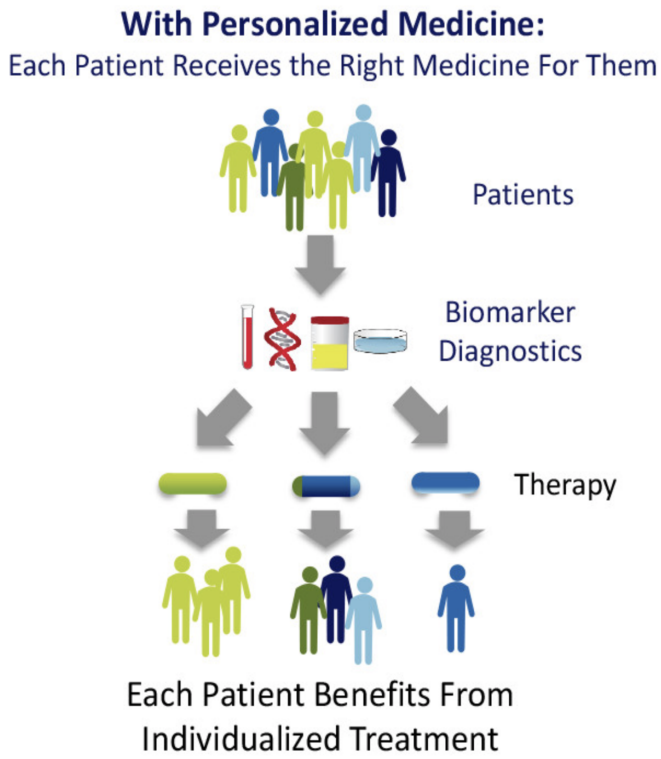
- Support for this presentation was provided by AbbVie. AbbVie participated in the review and approval of the content.

Subgroup Identification

- Goal: identify the baseline covariate profiles of patients who benefit from a experimental treatment, rather than determining whether or not the treatment has population level effect, can substantially 1) lessen the risk in undertaking a clinical trial, and 2) expose fewer patients to treatments that do not benefit them.
- Interest: Right treatment for a patient vs Right patient for a treatment.

Subgroup Identification





Subgroup Identification

Challenges to deal with:

- lack of power, since the sizes of subgroups are smaller than that of total study sample;
- multiplicity, due to the large number of subgroups examined;
- individualized treatment effects, due to heterogeneity among patients(genetics, environment, life style, etc);

OUTLINE

- Literature Review
- Model Specification
- Simulation Studies
- Real Trial Application
- Deep Learning Framework for Future Direction

Outline for section 1

Tree Methods

SIDES, IT, GUIDE, MOB, etc.

- SIDES: *Subgroup identification based on differential effect search (Lipkovich et al., 2011)*.
- IT: *Interaction trees, recursively partitions the data with splits chosen to optimize the objective function (Su et al.,2008)*.
- GUIDE: *Generalized unbiased interaction detection and estimation (Loh, 2002, 2009)*.
- MOB: *Model-based recursive partitioning (Seibold et al., 2016)*.

Non-tree Methods

SeqBT, PRIM, FindIt, OWE, etc.

- SeqBT: *Sequential bootstrapping and aggregating of threshold from trees (Huang et al.,2017)*.
- PRIM: *Patient rule induction method (G.Chen et al., 2015)*.
- FindIt: *Finding heterogeneous treatment effects (Imai and Ratkovic, 2013)*.
- OWE: *Outcome weighted estimation (S.Chen et al., 2017)*.

Motivation

Existing methods are limited to :

- mostly based on recursive partitioning, which may result in local optimum.
- difficult to handle multiple levels of categorical variables.
- testing the presence of treatment-covariate interactions rather than the resulting individualized treatment effects.
- lack multiplicity control.

Outline for section 2

Algorithm

The main algorithm can be summarized into 3 steps.

- 1. Exhaustive search of all subgroups up to 3 covariates.
- 2. Test the differential treatment effect between the subgroup identified by step 1 and the rest.
- 3. Using repeated cross-validation to control multiplicity.

Exhaustive Search

Suppose we have p covariates, each has K levels, given a pre-specified depth d, the algorithm will search over all possible combinations of covariates and their corresponding levels.
For example, suppose $X_1 \in \{0, 1, 2\}$ and $X_2 \in \{F, M\}$. Then the all possible combinations will be $(0, F), (0, M), (1, F), (1, M), (2, F), (2, M)$ and $(not0, F), (not0, M), (not1, F), (not1, M), (not2, F), (not2, M)$

- $d = 2$: $\binom{2}{p}(2K)^2 = O(p^2K^2)$
- $d = 3$: $\binom{3}{p}(2K)^3 = O(p^3K^3)$

Differential Treatment Effect Test

- ESa) let $\hat{\beta}^+$ and $\hat{\beta}^-$ be the estimated coefficient of treatment of the biomarker positive group and biomarker negative group. The test statistic is : $\frac{\hat{\beta}^+ - \hat{\beta}^-}{\sqrt{Var(\hat{\beta}^+ - \hat{\beta}^-)}}$, $\quad H_o : \beta^+ = \beta^-$

  Binary/Continuous response: logistic /linear regression
- ESb) For binary response, let $\hat{p}^+$ and $\hat{p}^-$ be the event-treatment rate, of biomarker positive group and biomarker negative group. The test statistics is:

  $\frac{\hat{p}^+ - \hat{p}^-}{\sqrt{Var(\hat{p}^+ - \hat{p}^-)}}$, $\quad H_o : p^+ = p^-$

  For continuous response, let $\bar{y_{tz}}, \bar{y_{rz}}, n_{tz}, n_{rz}$ are the mean responses and sample sizes in biomarker positive subgroup t and negative r group, $z = (0, 1)$, $\hat{\sigma}$ is pooled estimate of standard deviation. The test statistic is: $\frac{(\bar{y_{t1}} - \bar{y_{t0}}) - (\bar{y_{r1}} - \bar{y_{r0}})}{\hat{\sigma}\sqrt{n_{t0}^{-1} + n_{t1}^{-1} + n_{r0}^{-1} + n_{r1}^{-1}}}$, $\quad H_o : (y_{t1} - y_{t0}) - (y_{r1} - y_{r0}) = 0$

Data

Treatment arm	Endpoint	age<65	sex	marker1>0.2	marker1>0.5	marker1>0.7	BP	BMI
trt1	0	Y	M	Y	Y	Y	High	<20
trt1	0	Y	M	Y	Y	Y	High	<20
trt1	1	Y	F	Y	Y	N	High	20-28
trt1	1	N	F	Y	N	N	Normal	20-28
trt1	1	N	M	N	N	N	Normal	20-28
trt0	0	N	F	N	N	N	Normal	<20
trt0	0	N	F	Y	N	N	High	>28
trt0	0	Y	M	Y	Y	N	Normal	<20
trt0	1	Y	M	Y	Y	Y	High	>28

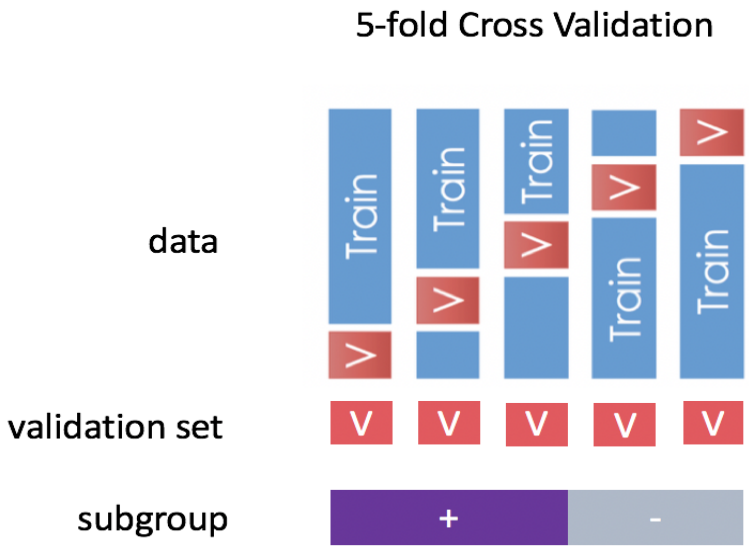
Training

subgroup of search depth 2	**Z-score**
Age < 65 and sex = M	2.38
Age !< 65 and biomarker1>0.2 = Y	1.95
sex = M and BP = high	0.35
...	...
BP != high and BMI < 20	**-2.55**
BP = normal and BMI !>2 8	0.97

Testing

samples	**subgroup**
sample1	Y
sample2	N
...	...
sample x	N

Repeated Cross Validation



- For each fold, use the best subgroup identified in the training set to label the patients in the validation set as biomarker positive or biomarker negative, combine all the validation set.
- Compare the treatment effect of group + and -, get the p-value(z score).
- Repeat CV procedure *n* times to get the median p-value(z score).

Outline for section 3

Simulation Model

Table 1: Models used for assessing the performance of subgroup identification methods

Model	Form	Treatment	Covariate	Error
A	$y_i = 1 + trt_i + I_{i1} + I_{i2} + \epsilon_i$	$Bernoulli(0.5)$	MVN(0,1)	N(0,1)
B_1	$y_i = 1 + trt_i + I_{i3} + I_{i4} + \gamma_{pred} * trt_i * I_{i1} * I_{i2} + \epsilon_i$	$Bernoulli(0.5)$	MVN(0,1)	N(0,1)
B_2	$y_i = 1 + trt_i - I_{i2} + I_{i3} + I_{i4} + \gamma_{pred} * trt_i * I_{i1} * I_{i2} + \epsilon_i$	$Bernoulli(0.5)$	MVN(0,1)	N(0,1)
B_3	$y_i = 1 + trt_i + I_{i1} + I_{i3} + I_{i4} + \gamma_{pred} * trt_i * I_{i1} + \epsilon_i$	$Bernoulli(0.5)$	Uniform$(-\sqrt{3},\sqrt{3})$	N(0,1)
B_4	$y_i = 1 + trt_i + I_{i3} + I_{i4} + \gamma_{pred} * trt_i * I_{i1} * I_{i2} + \epsilon_i$	$Bernoulli(0.5)$	MVN(0,1)	Exp(λ)
B_5	$y_i = \gamma_{pred} * (-1 + I_{i1} + I_{i2}) * trt_i + \epsilon_i$	$Bernoulli(0.5)$	MVN(0,1)	N(0,1)
B_6	$y_i = 1 + \gamma_{pred} * trt_i * I_{i1} * I_{i2} + \epsilon_i$	$Bernoulli(0.5)$	MVN(0,1)	N(0,1)

Simulation Model

Model A: is a null model without any treatment-covariate interaction.
Model B_1: is the main model to assess the treatment effect size λ_{pred}, power, partial recovery, full recovery under different sample sizes(n).
Model B_2: is the reverse model, there are some covariates in real situations are harmful for the response variable.
Model B_3: elucidates the effect of a different covariate distribution for different sample sizes(n) and treatment effect size λ_{pred}.
Model B_4: aim to evaluate the effect of different error distributions, the exponential family is a popular alternative of the normal distribution in health sciences.
Model B_5: the SIDES model.
Model B_6: a simple model without main effect, we utilize this model to evaluate the performance of subgroup identification methods when data come from failed but well-designed randomized clinical trials.

Simulation Setting

- Let y be the continuous response variable, we generate $x_1, ..., x_p$, $p = 20$ continuous/categorical covariates for each simulation model, and we are only interested in predictive effect in this study.
- $I_{i1} = I(x_{i1} > 0)$, $I_{i2} = I(x_{i2} > 0)$, $I_{i3} = I(x_{i3} > 0)$, $I_{i4} = I(x_{i4} > 0)$
- We consider treatment effect size $\lambda_{pred} = (0, 0.5, 1, 1.5, 2)$ and sample size $n = (250, 500, 750, 1000)$, covariates correlation $\rho = 0.2$.
- Each scenario is examined 500 simulation runs. In each simulation run, a new data set is generated and the methods selected are applied on this data set, significance level $\alpha = 0.05$.
- Our algorithm is written in R(ESa) and Java(ESb), for existing methods, the R packages can be found in http://biopharmnet.com/subgroup-analysis-software/. All simulations are done through iforge UIUC cluster.

Evaluation Metrics

- Power: the probability that the method can return subgroup with significant treatment effect.
- Partial recovery: the probability that the algorithm can identify at least one predictive covariate.
- Full recovery: the probability that the algorithm can only identify the predictive signatures we prespecified.
  We compare our proposed methods ESa, ESb with PRIM, SeqBT, and SIDES.

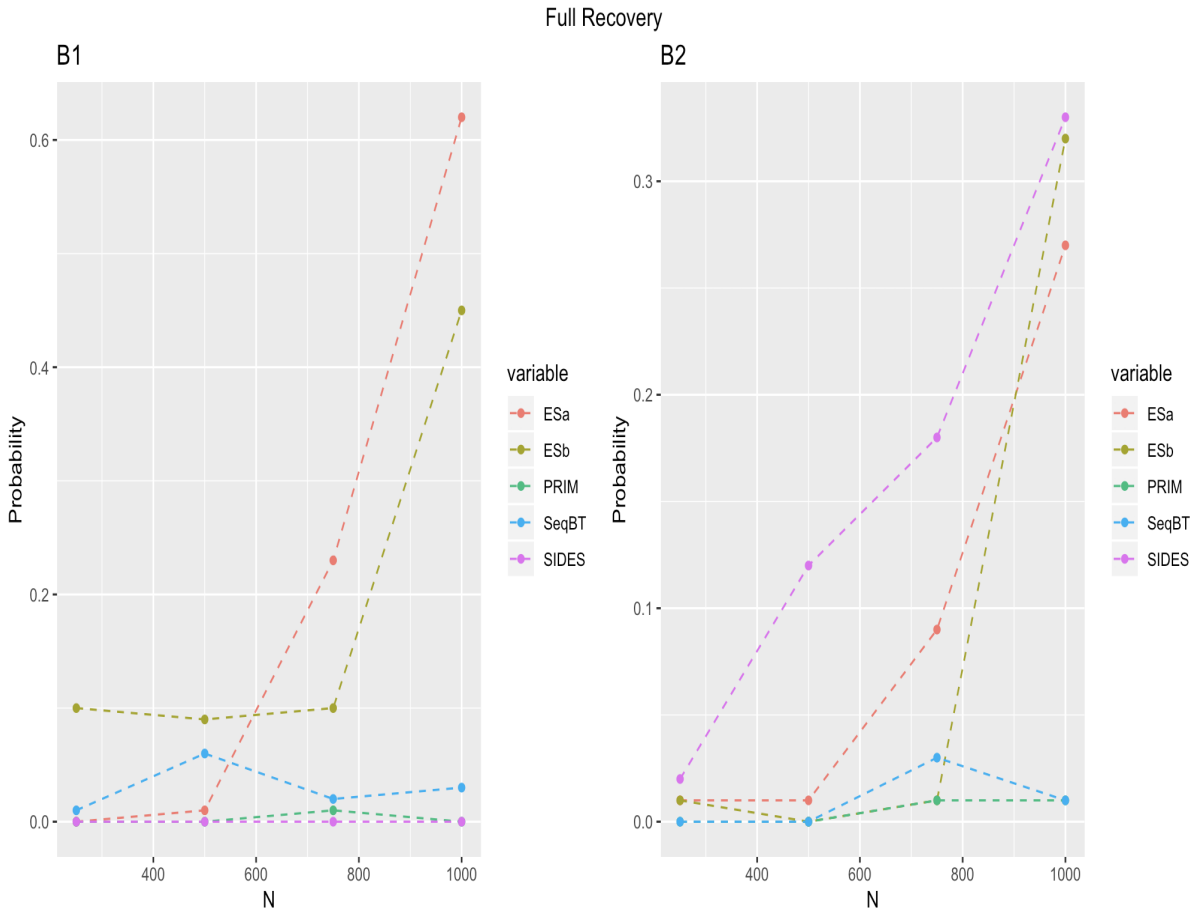
Result 1: Continuous Covariates FDR and Power

Table 2:Shows the simulation results of false discovery rate(type I error) and the power of identifying a significant treatment effect(B1 model).

N	λ	ESa	ESb	PRIM	Seq.BT	SIDES
250	0	0	0.07	0.01	0	0
500	0	0	0.01	0.02	0	0
750	0	0.01	0.05	0.04	0.03	0.01
1000	0	0.01	0.04	0.04	0.03	0.02
250	0.5	0	0.07	0.01	0.02	0.08
500	0.5	0	0.06	0.01	0.04	0.02
750	0.5	0.06	0.07	0.02	0.02	0.01
1000	0.5	0.06	0.04	0.06	0.11	0.12
250	1	0	0.28	0.05	0.13	0.19
500	1	0.03	0.28	0.08	0.29	0.26
750	1	0.46	0.38	0.34	0.38	0.68
1000	1	0.80	0.86	0.52	0.63	0.81
250	1.5	0.1	0.90	0.25	0.48	0.51
500	1.5	0.42	0.83	0.45	0.80	0.78
750	1.5	0.97	0.87	0.86	0.94	1
1000	1.5	1	1	0.97	1	1
250	2	0.3	1	0.45	0.81	0.91
500	2	0.75	1	0.76	0.99	1
750	2	0.99	1	0.93	1	1
1000	2	1	1	1	1	1

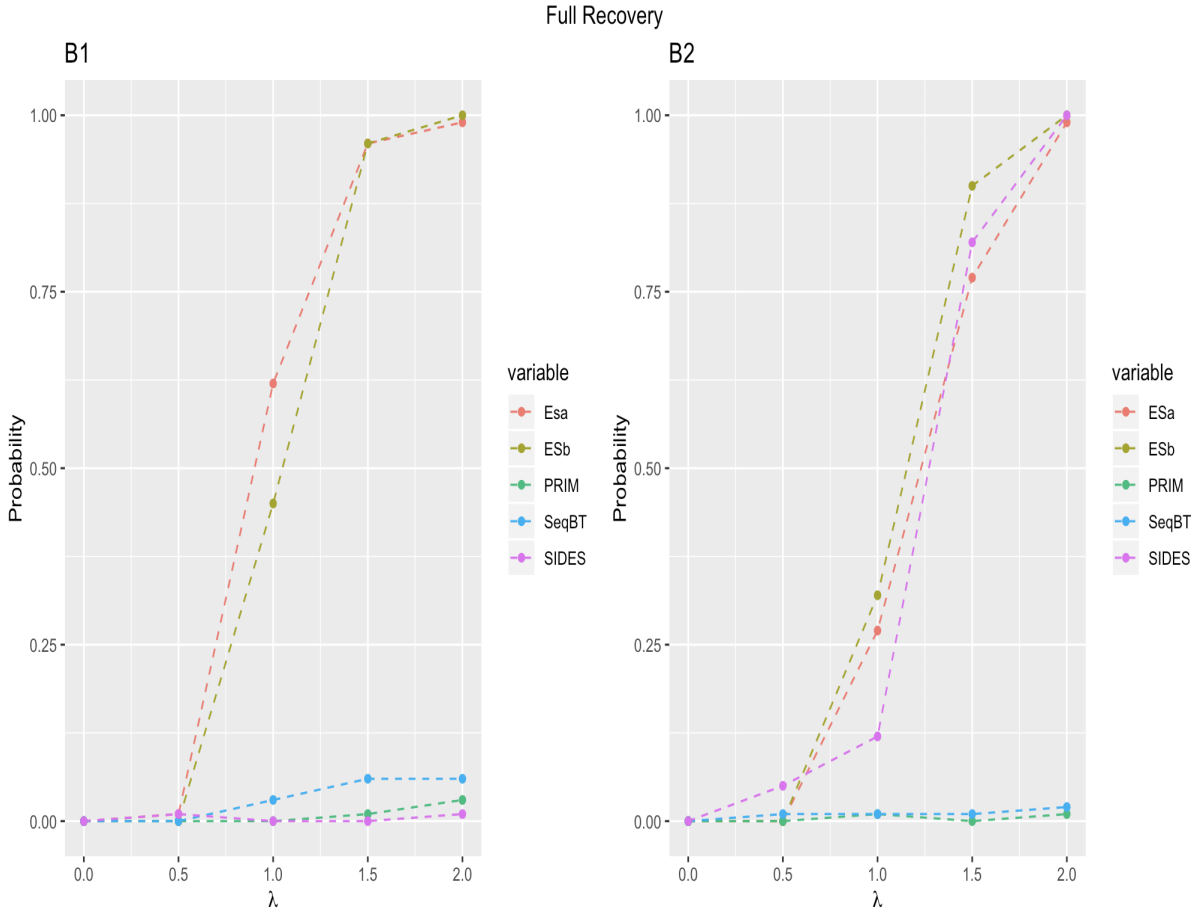
Result 1: Continous Covariates Full Recovery

Under different sample sizes(N) and $\lambda = 1$



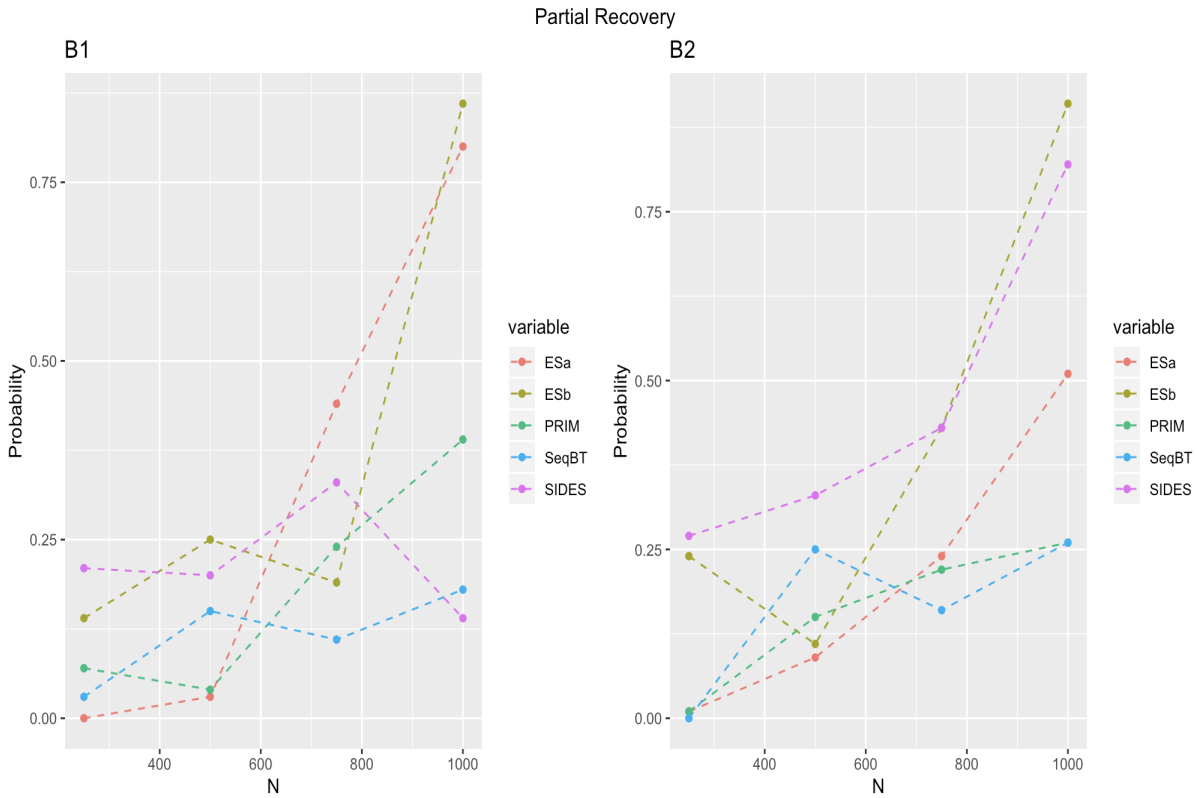
Result 1: Continous Covariates Full Recovery

Under different treatment effect size λ and sample size $N = 1000$



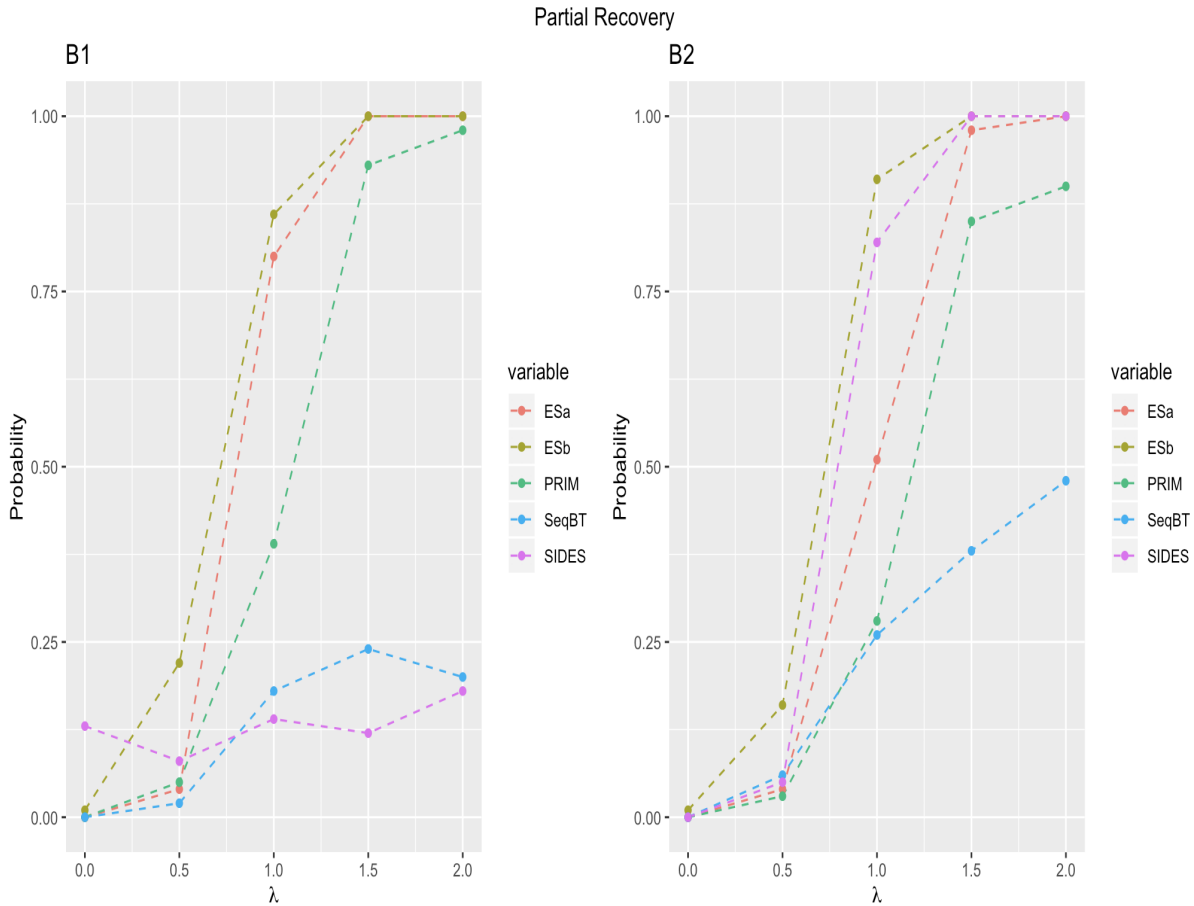
Result 1: Continous Covariates Partial Recovery

Under different sample sizes(N) and $\lambda = 1$



Result 1: Continous Covariates Partial Recovery

Under different treatment effect size λ and sample size $N = 1000$



Result 1: Summary

For continuous covariates:

- The proposed method has the lowest Type I error and the largest power of identifying subgroup with significant treatment effect.
- The proposed method has the best full and partial recovery rate in model B1, SIDES perform as good as our method in model B2 in terms of partial and full recovery rate.
- PRIM performs similarly as the proposed method only when treatment effect size $\lambda > 1.5$ for partial recovery rate.

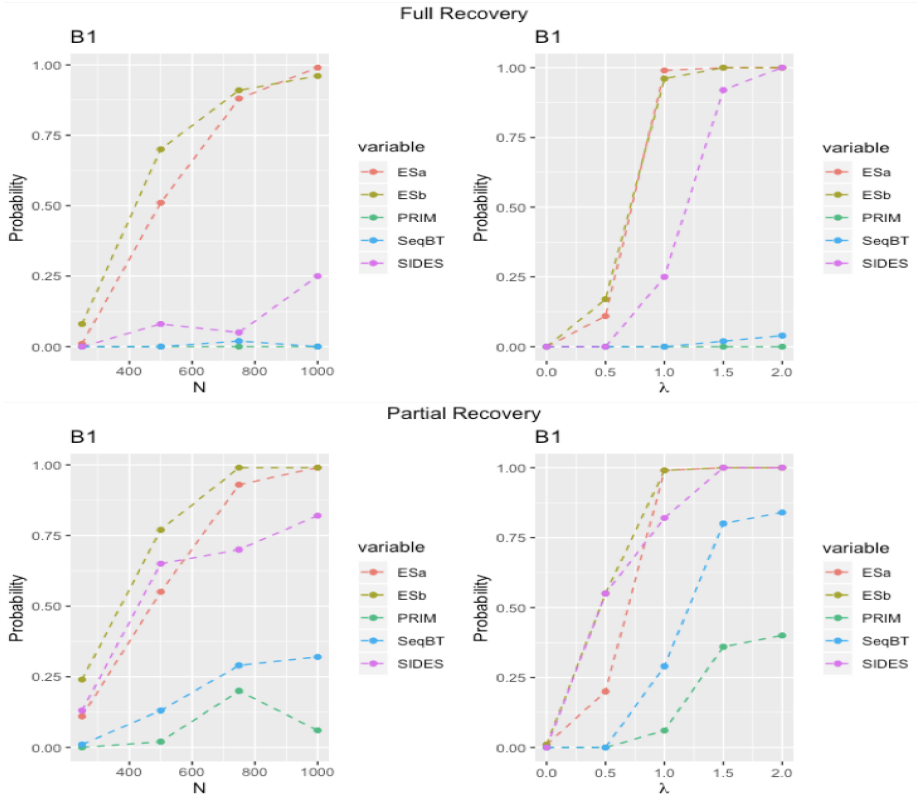
Result 2: Categorical Covariates FDR and Power

Table 3:Shows the simulation results of false discovery rate(type I error) and the power of identifying a significant treatment effect(B1 model).

N	λ	ESa	ESb	PRIM	Seq.BT	SIDES
250	0	0	0.04	0	0	0
500	0	0.03	0.07	0	0.02	0
750	0	0.01	0.06	0	0.02	0
1000	0	0.01	0.13	0	0.01	0
250	0.5	0.01	0.1	0	0.01	0.07
500	0.5	0.01	0.14	0.02	0.02	0
750	0.5	0.06	0.26	0.02	0.05	0.16
1000	0.5	0.20	0.56	0.01	0.01	0
250	1	0.1	0.25	0	0.03	0
500	1	0.56	0.79	0.02	0.18	0.08
750	1	0.94	0.99	0.27	0.51	0.05
1000	1	0.99	0.99	0.09	0.39	0.25
250	1.5	0.55	0.69	0.04	0.23	0
500	1.5	1	0.99	0.18	0.67	0.39
750	1.5	1	1	0.65	0.92	0.37
1000	1.5	1	1	0.46	0.99	0.92
250	2	0.92	0.97	0.19	0.68	0.17
500	2	1	1	0.28	0.95	0.67
750	2	1	1	0.85	1	0.84
1000	2	1	1	0.6	1	1

Result 2: Categorical Covariates Recovery

Results from model B1



Result 2: Summary

For categorical covariates:

- Overall, the proposed method has the largest power of identifying subgroup with significant treatment effect, but, ESb's Type I error(0.13) is a little higher compared to other methods.
- The proposed method has the largest full and partial recovery rate under different treatment effect sizes and sample sizes. For the full recovery rate, the proposed method is far better than the existing methods.
- SIDES has similarly partial recovery rate with the proposed method when sample size $N = 1000$.

Outline for section 4

1: The International Stroke Trial(IST)

IST is published non-AbbVie data (https://datashare.is.ed.ac.uk/handle/10283/128).
Goal: to establish whether early administration of aspirin, heparin, both or neither influenced the clinical course of acute ischemic stroke.

- 19,435 patients at 467 hospitals among 36 countries. Prospective, randomized, open treatment, and blined coutcome.
- Treatment: half on aspirin and half avoid aspirin; half on heparin and half avoid heparin.
- Primary endpoints(binary): death within 14 days; death or dependency at 6 months.

- We excluded patients whose initial events may not be acute ischemic stroke and noncompliant patients.
- Resulted dataset has 20 baseline covariates and 11,459 patients.
- After we get the subgroup we calculated the Z score to check the differential treatment effect.

Method	Subgroup	Z	#Obs	Time(s)
ESb	Age$\geq$47; Patrial=0	4.95	6679	35
PRIM	Patrial$\leq$0	3.07	9598	794
SeqBT	Age$>$ 48; Patrial$<$ 1;Rdelay$>$9	5.22	4619	3600
SIDES	Age$\geq$46	3.05	3077	261

IST: Real data simulation study

- We split the whole data set into different training and test set to run different methods. After we get the subgroup information in the training set, we label the test set patient as subgroup(+) and subgroup(-).
- For each split, we run 15 times to calculate the mean Z score and the computing time.

Method	25% Training	50% Training	75% Training	Time(s)
ESb	1.18	1.27	1.90	20
PRIM	0.90	0.95	1.17	610
SeqBT	0.91	0.99	1.13	3012
SIDES	NA	0.73	0.92	173

Real Application: Summary

- The proposed method ESb has the lowest computing cost. For IST data it only took 35s to finish run which is 1/100th compared to SeqBT.
- SIDES cannot find the subgroup under 25% Training data split.
- For IST data, PRIM always find subgroup with multiple covariates(>3) while the proposed method is quite stable.

Discussion

- The proposed method performs favorably than the existing methods in terms of full and partial biomarker recovery rate for both continuous and categorical covariates.
- The proposed method can better handle categorical covariates with multiple levels and will return exactly the categorical level, rather than SIDES and SubgrpID can only output subgroups like like $X < c$ or $X > c$.
- For mixed types of continuous and categorical covariates, the proposed method is more powerful(IST data).

Outline for section 5

Future directions

- High-dimensional and Non-linear complex relationship between variables and outcome.
- Leverage the available side information to increase how many significant outcomes we can detect.
- Multiple treatments.

Deep Learning Framework

We model the test statistic as arising from a mixture model of two components, the null (f_0) and the alternative (f_1). An experiment-specific weight c_i then models the prior probability of the test statistic coming from the alternative (the probability of the treatment having an effect). We place a beta prior on each experiment-specific prior c_i and model the parameters of the hyperprior with a black box function G parameterized by θ; in our implementation, G is a deep neural network. The complete model is:

$$\begin{aligned} z_i &\sim h_i f_1(z_i) + (1 - h_i) f_0(z_i) \\ h_i &\sim \text{Bernoulli}(c_i) \\ c_i &\sim \text{Beta}(a_i, b_i) \\ a_i, b_i &= G_{\theta,i}(X)\,. \end{aligned} \tag{1}$$

Then optimize θ by integrating out h_i and maximizing the complete data log-likelihood,

We fit the model in (1) with stochastic gradient descent (SGD) on an L_2-regularized loss function,

$$\underset{\theta \in |\theta|}{\text{minimize}} \quad -\sum_i \log p_\theta(z_i) + \lambda \parallel G_\theta(X) \parallel_F^2 , \tag{3}$$

where $\| \cdot \|_F$ is the Frobenius norm. For computational purposes, we approximate the integral in (2) by a fine-grained numerical grid. Once the optimized parameters $\hat{\theta}$ are chosen, we calculate the posterior probability of each test statistic coming from the alternative,

$$\begin{aligned} \hat{w}_i &= p_{\hat{\theta}}(h_i = 1 | z_i) \\ &= \int_0^1 \frac{c_i f_1(z_i) \text{Beta}(c_i | G_{\hat{\theta},i}(X))}{c_i f_1(z_i) + (1 - c_i) f_0(z_i)} dc_i . \end{aligned} \tag{4}$$

Assuming the posteriors are accurate, rejecting the i^{th} hypothesis will produce $1 - \hat{w}_i$ false positives in expectation. Therefore we can maximize

the total number of discoveries by a simple step down procedure. First, we sort the posteriors in descending order by the likelihood of the test statistics being drawn from the alternative. We then reject the first q hypotheses, where $0 \leq q \leq n$ is the largest possible index such that the expected proportion of false discoveries is below the FDR threshold. Formally, this procedure solves the optimization problem,

$$\begin{aligned} \underset{q}{\text{maximize}} \quad & q \\ \text{subject to} \quad & \frac{\sum_{i=1}^{q}(1-\hat{w}_i)}{q} \leq \alpha\,, \end{aligned} \tag{5}$$

for a given FDR threshold α. By convention $\frac{0}{0} = 0$(Wesley et al., 2018).

Acknowledgement

- Hong Zhang
- Yan Sun, Xin Huang, Yingtao Bi, Feng Hong, James Fann, Gyan P Srivastava
- Ivan Chan

References I

1. Lipkovich, I., Dmitrienko, A., Denne, J., Enas, G. (2011). Subgroup identification based on differential effect searcha recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine, 30, 2601-2621.*

2. Su, X., Zhou, T., Yan, X., Fan, J., Yang, S. (2008). Interaction trees with censored survival data. *International Journal of Biostatistics, 4.*

3. Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica, 12, 361-386.*

4. Loh, W.-Y. (2009). Improving the precision of classification trees. *Annals of Applied Statistics, 3, 1710-1737.*

References II

5. Seibold, H., Zeileis, A., Hothorn, T. (2016). Model-based recursive partitioning for subgroup analyses. *International Journal of Biostatistics, 12, 45-63.*

6. Huang, X., Sun, Y., Trow, P., Chatterjee, S., Chakravartty, A., Tian, L., Devanarayan, V. (2017). *Patient subgroup identification for clinical drug development. Statistics in Medicine, 36, 1414-1428.*

7. Chen, G., Zhong, H., Belousov, A., Devanarayan, V. (2015). A PRIM approach to predictive-signature development for patient stratification. *Statistics in Medicine, 34, 317-342.*

References III

8. Imai, K., Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics, 7, 443-470.*

9. Chen, S., Tian, L., Cai, T., Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics, 73, 199-1209.*

10. Tansey,W, Wang, YX, Blei,D, Rabadan R (2018). Black Box FDR. *ICML, 80.*