NOTE FOR SUBGROUP IDENTIFICATION

Xiaowei Yin

1 Summary

Identification of subgroups that have heterogeneous treatment effect is a concern for different fields. Here researchers are interested in modeling the relationship between outcome Y, treatment Z, covariate X and the subgroup indicator S. Usually, it can be seen as a special case of model selection. Various models and selection methods are proposed and many include treatment-covariate interaction or treatment-split interaction or covariate-split(subgroup indicator) interaction to model treatment heterogeneity and discover promising subgroup. Then it is typically that a variable selection step is employed to reduce model multiplicity and improve robustness. The discovered subgroup is then estimated using debiased methods for confirmation and statistical inference.

It's beneficial to first introduce the classification scheme in this field for an overview. The existing literature can be classified in different aspects and I listed them in (section1.1 & section 1.2). Apart from that, I also found that there are some common concerns like selection bias and multiplicity control in this field that worth mentioning in advance (section 1.3 & section 1.4).(Lipkovich et al., 2017)(Loh et al., 2019)

1.1 classification scheme

Framework for personalized medicine:

- Identifying the right patient for a given treatment (usually on a treatment that provides minimal or no benefit in the overall population)(quantitative interaction)
- Identifying the right treatment for a patient. (find OTR or policy for a given subpopulation)(qualitative interaction)

Focused problems:
Subgroup identification can be roughly classified into confirmatory subgroup analysis and exploratory subgroup analysis. And for subgroup discovery, the methods can be further classified into the following schemes depending on their different focus.

- Global outcome modeling methods
- Global treatment effect modeling methods

- Optimal treatment regimes
- Local modeling methods

used methods:

- tree-based method
- non-tree method

Key features:
Here are some features that may occur in a method related to subgroup identification. It can help us evaluate a method from the following aspect.

- modeling type:
  Freq (Frequentist), Bayes (Bayesian);
  P (parametric), SP (semiparametric), NP (nonparametric);
- dimensionality of the covariate space:
  low, medium, high;
- results produced by the method:
  B (selected biomarkers or biomarker ranking based on VI scores that can be used for tailoring), P (predictive scores for individual treatment effects), T (optimal treatment assignment), S (identified subgroups);
- evaluation of the Type I error rate/false discovery rate for the entire subgroup search strategy
- application of complexity control to prevent data overfitting
- control (reduction) of selection bias when evaluating candidate subgroups
- Availability of 'honest' estimates of treatment effects in identified subgroups

This classification of available methods provides some insight as to the situations when different methods may be particularly applicable. For example, methods that evaluate optimal regimes are useful in large Phase III or IV trials that compare several active treatments in a diverse population. Methods that utilize penalized regression and ensemble learning can handle very large sets of candidate covariates. As a consequence, these methods can be used in settings where the sample size is rather small, including early-stage trials, and the main focus is on selecting biomarkers rather than specific patient subgroups that can be utilized in subsequent Phase III trials. Tree-based methods are useful when there are a few candidate biomarkers, for example, 15–20 biomarkers, in relatively large datasets (say, with 1000–2000 patients) and subgroups can be reliably estimated. Evaluation of biomarkers using Bayesian shrinkage regression models such as models studied in is well suited to evaluating post-hoc hypotheses or meta-analysis with a relatively small number of subgroups defined by units where the exchangeability assumption is reasonable. Examples include studies that focus on the effect of multiple countries or demographic groups.

1.2 General comparison of subgroup identification methods

The subgroup identification method, given the amount of uncertainty and lack of knowledge about subpopulations of patients who may experience enhanced treatment effect. Nonparametric method (e.g. recursive partitioning) appear more flexible and efficient compared with parametric approaches in that they support subgroup exploration within a very broad 'model space'

Further, unlike standard recursive partitioning methods (e.g., CART) that aim at identifying subgroups with heterogeneous outcome values, partitioning methods for personalized medicine rely on a variety of splitting criteria that are modified appropriately to focus on subgroups with a differential treatment effect. This is typically achieved by incorporating treatment-by-splitting (e.g. IT) covariate interaction effects.

eg: compare univariate regression and tree-based regression in simulation experiment

1. Univariate regression approach:
- They ignore potential synergistic effects of two or more biomarkers by failing to account for higher-order interaction effects.

2. Tree-based regression models:
- This model run a patient down the tree and the predicted value is defined as the average outcome with in the resulting terminal node.
- Problems of tree-based model are that: The treatment variable is not comparable with the strong prognostic biomarkers and the tree-fitting process wrongly selected subgroups with differential outcomes rather than differential treatment effect (predictive variables) that we care about.

1.3 Multiplicity adjustment and complexity control

- **Complexity control** : Biomarker/subgroup identification can be considered as a special case of model selection. And then the idea of complexity control can be relate to the trade-off between bias and variance in ML.
- **Multiplicity adjustment** : weak control of the probability of incorrect subgroup selection associated with a subgroup identification strategy can be implemented based on resampling methods.

Multiplicity -adjusted treatment effect p-value:

As an example, performing a greedy search for subgroups by brute force, that is, by a complete enumeration of all possible subgroups that can be formed by, say, up to three biomarkers, is likely to generate spurious subgroups with highly significant treatment effect p-values. However, the probability of observing a similar significant treatment effect within these subgroups in another study will be low. //Replicating the entire strategy on the reference (null) data is likely to also generate subgroups with highly significant p-values. Therefore, resampling-based multiplicity-adjusted p-values (i.e., the proportion of null sets with p-values as small as or smaller than the observed p-value)

would be relatively large. The larger the multiplicity-adjusted treatment effect p-value is, the more unrepresentative the significance in p value is, so it should be adjusted more unsignificant(larger) in compromise for its unrepresentativeness. This phenomenon can be understood as a large variance.

More generally, multiplicity adjustments ought to be used in combination with controlling the complexity of the subgroup selection process. Performing an unconstrained search for subgroups followed by a multiplicity adjustment may be an inefficient strategy because:

• It may result in identifying patient subgroups that have a low chance of being replicated in an independent dataset.
• The resulting multiplicity adjustment may be too conservative, which will lead to very large multiplicity-adjusted treatment effect p-values within the selected subgroups.

To solve the above problems, less greedy strategies that put an appropriate 'constraint jacket' on model space may be employed to result in less complex subgroups based on fewer biomarkers. The reduced models space results in a lower multiplicity burden and therefore a smaller multiplicity penalty when computing multiplicity-adjusted p-values. As a consequence, a modestly significant observed p-value associated with a subgroup based on constrained subgroup search is likely to translate into a much smaller adjusted p-value compared with that obtained after unconstrained search.

Several approaches have been proposed recently to avoid 'greediness' and overfitting in subgroup search:

- **frequentist methods**
  employing complexity penalties, typically determined by resampling-based methods, for example, methods based on penalized regression [RowSi,FindIt,] and tree-based methods [QUINT];
- **ensemble learning methods**
  that average over a large number of 'learners' to shrink the contribution of noise covariates to zero [VT, SIDEScreen,];
- shrinkage and model averaging via Bayesian methods;
- methods that use 'indirect' or less direct criteria for variable/subgroup selection that avoid exhaustive search for subgroups with desired features. [Guide];

1.4 Bias-corrected treatment effect estimates

1. bias:
One of the most challenging tasks in subgroup identification is obtaining unbiased and reliable estimates of treatment effects in the selected patient subgroups.

In order to obtain unbiased estimates, we noormally requires:

- **honest estimation:** additional independent (or test) data. When no test datasets are available.[Causal Trees]

- **resampling method:** bootstrap or CV

When resampling data have been used for tuning a method's complexity parameters, the same data cannot be re-used to compute 'honest' estimates of treatment effect. As a general principle, when using resampling methods for computing bias-corrected subgroup effects, it is important that the entire search strategy (including estimation of any data-driven tuning parameters) be implemented afresh on each dataset.

2. treatment effect estimation:

- **expected treatment effect** in a specific subgroup and its excess over that in the overall population;
- **utility function** evaluated on a subgroup that takes into account the 'treatment burden' based on safety and/or extra costs that may also reflect the minimal clinically meaningful treatment effect in the subgroup;
- **power or predictive power** of a future trial where the identified subgroup will be used as part of a tailoring strategy, for example, the trial may utilize an enrichment design based on this patient subgroup;
- **value function** of the optimal treatment assignment rule based on the identified biomarkers/subgroups compared with a rule that assigns all patients to the same treatment.

2 Global outcome modeling

Methods fall in this regime model the relationship between the outcome and covariate, so the main feature of this class of method is that their response variable is the observed/potential outcome.

2.1 Findit

Summary: FindIt (Imai and Ratkovic, 2013) used penalized regression to select predictive variables and the penalty serve as a multiplicity control. Especially, this method uses 2 distinct penalty term for prognostic variables and predictive variables respectively, aiming to dress the problem that predictive variables are always weaker compared with prognostic variables.

Model:
penalized regression is used to select variables:

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} L\left(y_i, f\left(\mathbf{x}_i \mid \boldsymbol{\beta}\right)\right) + J_{\lambda}(\boldsymbol{\beta}) \right)$$

Penalty function is the lasso penalty (l1 penalty), but two separate penalty parameters are introduced for prognostic and predictive variables. Penalty parameter was defined through GCV.

$$J_\lambda(\boldsymbol{\beta}) = \lambda_{ul} \sum_{j=1}^{p_z} \left|\beta_j^{(u)}\right| + \lambda_v \sum_{j=1}^{p_v} \left|\beta_j^{(v)}\right|$$

Estimated Treatment contrast was calculated through the estimated outcome.

$$\hat{z}_i = \frac{1}{2}\left[\hat{f}_{tr}\left(\mathbf{x}_i, 1\right) - \hat{f}_{tr}\left(\mathbf{x}_i, 0\right)\right]$$

note: The penalty term can serve as a multiplicity control, that is the l1 penalty shrink some irrelevant covariate's coefficient to zero (variable selection).

2.2 VirtualTwins

Summary: VirtualTwins (Foster et al., 2011) is a two-stage method. It firstly models the outcome using random forest and compute treatment contrast on the estimator of the outcome, then it uses the computed treatment contrast to find subgroup.

Model:
stage1: response(f) is estimated using random forest (in the potential outcome framework) and treatment contrast(z) is computed.

$$z_i = \hat{f}\left(\mathbf{x}_i, 1\right) - \hat{f}\left(\mathbf{x}_i, 0\right)$$

$$z_i = \operatorname{logit} \hat{f}\left(\mathbf{x}_i, 1\right) - \operatorname{logit} \hat{f}\left(\mathbf{x}_i, 0\right)$$

Stage2: the estimated contrasts are used as observed values for growing a regression tree. The tree is pruned using CV for a multiplicity control. Then we can use the fitted tree to obtain predictions.

$$\hat{z}(R) = \frac{1}{|R|} \sum_{j \in R} z_j$$

Subgroup identification criteria:
Then each terminal node is classified into one of the two outcome groups based on the 'majority vote' within the node, and the enhanced subgroup is determined with the estimator greater than a prespecified constant (clinically important threshold).

$$\widehat{u}(R) = I\left(\frac{1}{|R|} \sum_{j \in R} u_j \geqslant 0.5\right)$$

Measurement of treatment effect heterogeneity:
treatment benefit is defined to estimate the treatment effect of the selected subgroup. It is defined as the 'excess' treatment effect in the true subgroup S over the overall population effect:

$$Q(S) = \{E(f(\mathbf{X}, 1) \mid \mathbf{X} \in S) - E(f(\mathbf{X}, 0) \mid \mathbf{X} \in S)\} - \{E(f(\mathbf{X}, 1)) - E(f(\mathbf{X}, 0))\}$$

(should be evaluate on independent dataset to get unbiased estimator) Because the true subgroup S is unknown, the treatment benefit needs to be evaluated for the estimated subgroup $\hat{S}$

Model based estimate:

$$\widehat{Q}(\widehat{S}) = \left(\frac{1}{|\widehat{S}|} \sum_{i:x_i \in \widehat{S}} \widehat{f}(\mathbf{x}_i, 1) - \frac{1}{|\widehat{S}|} \sum_{i:x_i \in \widehat{S}} \widehat{f}(\mathbf{x}_i, 0) \right) - \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{f}(\mathbf{x}_i, 1) - \frac{1}{N} \sum_{i=1}^{N} \widehat{f}(\mathbf{x}_i, 0) \right)$$

Data based estimate:

$$\hat{Q}(\hat{S}) = \left(\frac{1}{\left|\hat{S}_1\right|} \sum_{i:x_i \in \hat{S}_1} y_i - \frac{1}{\left|\hat{S}_0\right|} \sum_{i:x_i \in \hat{S}_0} y_i \right) - \left(\frac{1}{N_1} \sum_{i:t=1} y_i - \frac{1}{N_0} \sum_{i:t=0} y_i \right)$$

Intermediate estimate:

$$\widehat{Q}(\widehat{S}) = \left(\frac{1}{\left|\widehat{S}_1\right|} \sum_{i:x_i \in \hat{S}_1} \hat{f}(\mathbf{x}_i, 1) - \frac{1}{\left|\widehat{S}_0\right|} \sum_{i:x_i \in \hat{S}_0} \hat{f}(\mathbf{x}_i, 0) \right) - \left(\frac{1}{N_1} \sum_{i:t_i=1} \hat{f}(\mathbf{x}_i, 1) - \frac{1}{N_0} \sum_{i:t_i=0} \hat{f}(\mathbf{x}_i, 0) \right)$$

0.632 estimator balances the re-substitution and 'out-of-bag' estimators.

bias control: The estimated treatment benefit may be overoptimistic. A non-parametric bootstrap was employed to create bias-corrected estimate:

$$O_b = \hat{Q}_b\left(\hat{S}_b\right) - \widehat{Q}\left(\hat{S}_b\right)$$

$$\widehat{\text{Bias}} = B^{-1} \sum_{b=1}^{B} O_b$$

$$\widehat{Q}_{\text{cor}}(\hat{S}) = \widehat{Q}(\hat{S}) - \widehat{\text{Bias}}$$

2.3 Logistic-Normal mixture model

Summary: The method (Shen and He, 2015) adopted the follows the idea of mixture of experts, and takes the form of Logistic-Normal mixture model to model the subgroup heterogeneous treatment effect. And EM algorithm is employed to solve the problem in maximizing the likelihood.

Model:
employed Logistic-Normal mixture model to model the subgroup heterogeneous treatment effect.

$$Y_i = \boldsymbol{Z}_i^T (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 \delta_i) + \varepsilon_i$$

$$P(\delta_i = 1 \mid \boldsymbol{X}_i, \boldsymbol{Z}_i) = \pi\left(\boldsymbol{X}_i^T \boldsymbol{\gamma}\right) \equiv \exp\left(\boldsymbol{X}_i^T \boldsymbol{\gamma}\right) / \left(1 + \exp\left(\boldsymbol{X}_i^T \boldsymbol{\gamma}\right)\right)$$

$$P\left(\delta_i = 0 \mid \boldsymbol{X}_i, \boldsymbol{Z}_i\right) = 1 - P\left(\delta_i = 1 \mid \boldsymbol{X}_i\right)$$

Likelihood ratio test: a likelihood ratio test is introduced in testing whether there exists predictive subgroups for different treatment effects. This test can be used in confirmatory subgroup identification as well as exploratory subgroup identification.

$$\begin{aligned} f(Y, \boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\eta}) \propto \pi\left(\boldsymbol{X}^T \boldsymbol{\gamma}\right) \varphi_\sigma\left(Y - \boldsymbol{Z}^T\left(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2\right)\right) \\ + \left(1 - \pi\left(\boldsymbol{X}^T \boldsymbol{\gamma}\right)\right) \varphi_\sigma\left(Y - \boldsymbol{Z}^T \boldsymbol{\beta}_1\right) \end{aligned}$$

EM test: since the null model of no-subgroups is not an interior point in the alternative space, therefore the likelihood ratio test do not have a good chi-square limiting distribution, and also the likelihood involves two parameters that increase the difficulty in maximizing the likelihood.

To address the above problem, the author of this paper proposed EM test. They firstly find that the test statistic has a good chi-square limit distribution with a fixed gamma. Then considering the fact that we have no prior knowledge of gamma, they adopted the EM algorithm to adaptively update the estimation of gamma.

E-step:

$$a_i^{(k)} = P\left(\delta_i = 1 \mid Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i; \eta^{(k)}\right), i = 1, \ldots, n$$

M-step:

$$\boldsymbol{\theta}^{(k+1)} = \underset{\theta}{\operatorname{argmax}} \left(\sum_{i=1}^{n} \left[a_i^{(k)} \log f\left(Y_i \mid \delta_i = 1, \boldsymbol{Z}_i; \boldsymbol{\theta}\right) \; + \left(1 - a_i^{(k)}\right) \log f\left(Y_i \mid \delta_i = 0, \boldsymbol{Z}_i; \boldsymbol{\theta}\right) \right] \right)$$

$$\boldsymbol{\gamma}_{\text{temp}}^{(k+1)} = \underset{\gamma}{\operatorname{argmax}} \left(\sum_{i=1}^{n} \left[a_i^{(k)} \log P\left(\delta_i = 1 \mid \boldsymbol{X}_i; \boldsymbol{\gamma}\right) \; + \left(1 - a_i^{(k)}\right) \log P\left(\delta_i = 0 \mid \boldsymbol{X}_i; \boldsymbol{\gamma}\right) \right] \right)$$

They find that the proposed EM test statistic need small iteration to achieve a good asymptotic limit distribution property.

3 Global treatment effect modeling

Instead of modeling the outcome, the methods that falls into this area choose to model the treatment effect directly. This may bring some benefit by avoiding the disturbance of prognostic variables.

3.1 IT

Summary: Interactive Tree (Su et al., 2008) includes the treatment-by-split-interaction term to focus on splits that make the resulting treatment contrast in terminal nodes differs

more. Employed test statistic to test the null hypothesis that coefficient of treatment-by split interaction term=0.

Model:
Tree growing step: For each parent node, the split s^* is selected to maximize $G(s)$ over all allowable splits for all candidate covariates. Here $G(s)$ is the likelihood-ratio test statistic for the following hypothesis problem:

$$\text{H0: } h\left(u \mid t_i, s_i\right) = h_0(u) \exp\left(b_1 s_i + b_2 t_i\right)$$

$$\text{H1: } h\left(u \mid t_i, s_i\right) = h_0(u) \exp\left(a_1 s_i + a_2 t_i + a_3 s_i t_i\right)$$

LRT statistic: (represent the treatment contrast)

$$G(s) = -2\left(l_2 - l_1\right)$$

Note: this step find split that maximize the treatment contrast.

Steps 2 and 3: tree pruning and estimation.

multiplicity control: interaction-complexity criterion for a tree structure is introduced for pruning a tree.

$$G_a(\mathcal{T}) = G(\mathcal{T}) - \alpha\left(|\mathcal{T}_{\text{term}}| - 1\right)$$

$$G(\mathcal{T}) = \sum_{s \in \mathcal{T} - \mathcal{T}_{\text{yerm}}} G(s)$$

$G(\mathcal{T})$ is the amount of treatment heterogeneity associated with all the splits within a given tree structure.

bias control: a bias-corrected estimate $\hat{G}(\mathcal{T}_i)$ was obtained using a resampling-base method. Because G will be overoptimistic when computed by resubstitution.

Subgroup identification criteria: Select final sub-tree maximize the following criterion:

$$\widehat{G}_{\alpha_c}\left(\mathcal{T}_i\right) = \widehat{G}\left(\mathcal{T}_i\right) - \alpha_c\left(|\mathcal{T}_{\text{term},i}| - 1\right)$$

The terminal nodes in the selected tree indicates the identified subgroup that has most heterogeneous treatment effect.

3.2 GUIDE

Summary: GUIDE (Loh et al., 2014) is a 2 stage tree-based selection procedure that first choose covariate x by chi-squared test, then decide the optimal cutoff. And it has a non-greedy nature that the estimated treatment effect is unbiased.

3.3 MOB

Summary: MOB(Model-based recursive partitioning) (Seibold et al., 2016) view subgroup identification as a problem of model segmentation. Here the subgroups are the

submodels that have different coefficient of the prognostic and predictive variable. A hypothesis test is carried out to test the independence of partial score (parameter instability) and partitioning variables. Variables that reject the null hypothesis are used to define subgroup in a recursive partitioning pattern.

Model:
The treatment effect can be modeled as a function of patient characteristics. And the subgroups can be identified by the different coefficient in each subgroup. And we can get an estimation of the coefficients through minimizing negative log-likelihood:

$$\hat{\vartheta} = \arg\min_{\vartheta} \sum_{i=1}^{N} \Psi\left((y, x)_i, \vartheta\right)$$

The patient subgroups can be defined as a partition $\{\mathcal{B}_b\}\,(b = 1, ..., B)$ with partitioning variables Z. The subgroupspecific model parameters are then $\vartheta(b)$ with $\vartheta(b) = (\alpha(b), \beta(b), \gamma, \sigma)^\top$

$$(\hat{\vartheta}(b))_{b=1,...,B} = \arg\min_{\vartheta(b)} \sum_{i=1}^{N} \sum_{b=1}^{B} \mathbb{I}\left(\mathbf{z}_i \in \mathcal{B}_b\right) \Psi\left((y, \mathbf{x})_i, \vartheta(b)\right)$$

$$\alpha(\mathbf{z}) = \sum_{b=1}^{B} \mathbb{I}\left(\mathbf{z} \in \mathcal{B}_b\right) \cdot \alpha(b) \quad \text{and} \quad \beta(\mathbf{z}) = \sum_{b=1}^{B} \mathbb{I}\left(\mathbf{z} \in \mathcal{B}_b\right) \cdot \beta(b)$$

the score function ψ is introduced to quantify coefficient instability, i.e. the gradient of the objective function Ψ with respect to non-constant intercept $\alpha(b)$ and treatment effects $\beta(b)$.

$$\psi_\alpha((Y, \mathbf{X}), \vartheta) = \partial\Psi((Y, \mathbf{X}), \vartheta)/\partial\alpha \text{ and } \psi_\beta((Y, \mathbf{X}), \vartheta) = \partial\Psi((Y, \mathbf{X}), \vartheta)/\partial\beta$$

In order to formally detect deviations from independence between the partial score functions and the partitioning variables, model-based recursive partitioning utilises independence tests. The null hypotheses are as follows:

$$\begin{aligned} H_0^{\alpha,j} &: \psi_\alpha((Y, \mathbf{X}), \hat{\vartheta}) \perp Z_j, j = 1, \ldots, J \\ &\quad \text{and} \\ H_0^{\beta,j} &: \psi_\beta((Y, \mathbf{X}), \hat{\vartheta}) \perp Z_j, j = 1, \ldots, J \end{aligned}$$

model-based recursive partitioning selects the partitioning variable Z_j^* associated with the highest correlation to any of the partial score functions (smallest p-value). The procedure of testing independence of partial score functions and partitioning variables is repeated recursively until deviations from independence can no longer be detected. The resulted leaf nodes contain the patients of the different subgroups and specify the partition-specific models.

3.4 CT

Summary: Causal Trees (Wager and Athey, 2018) focus on the unbiasd estimation of treatment effect. And the asymptotic property of random forest that grown following the proposed procedure for growing a causal tree has been proofed.

Double-Sample Trees: Double-sample trees split the available training data into two parts: one half for estimating the desired response inside each leaf, and another half for placing splits.

Propensity Trees: Propensity trees use only the treatment assignment indicator to place splits, and save the responses Y_i for estimating treatment effect.

These 2 procedures can help build honest trees that avoid the overoptimistic bias.

3.5 QUINT

Summary: QUINT(Qualitative interaction trees) (Dusseldorp and Van Mechelen, 2014) looks for "qualitative interactions," where one treatment performs better than another in one subgroup and worse in another subgroup.

4 Modeling optimal treatment regimes (OTR)

The methods fall into this region emphasize identifying an optimal treatment for a given patient. The model can be either modeling the global outcome or the treatment contrast.

general model:
1. Treatment regime (individual treatment rule) d(X) is defined as the function that maps a patient's covariate X to treatments.
2. Potential outcome associated with a specific regime d(X):

$$\tilde{Y}(d(\mathbf{X})) = \tilde{Y}(1)d(\mathbf{X}) + \tilde{Y}(0)(1 - d(\mathbf{X}))$$

3. Value function that model the expected rewards if all patients follows the rule d(x)

$$V[d(\mathbf{X})] = E[\tilde{Y}(d(\mathbf{X}))]$$

4. An OTR is defined as the optimal treatment allocation strategy d(X) that maximize the expected rewards.

$$d_{\text{opt}}\ (\mathbf{X}) = \underset{d}{\text{argmax}} V[d(\mathbf{X})]$$

For example, if the outcome function f(X,T) is known, then we can have:

$$\widehat{d}_{\text{opt}}\ (\mathbf{X}) = I(\widehat{f}(\mathbf{X}, 1) > \widehat{f}(\mathbf{X}, 0))$$

Further, if outcome function is linear, then an optimal individual treatment rule depends on the covariates only through z(x).

$$\widehat{d}_{opt}(\mathbf{X}) = I(\widehat{z}(\mathbf{X}) > 0)$$

4.1 OWL

Summary: The OWL (outcome weighted learning) (Fu et al., 2016) is the method that follows the general method of modeling optimal treatment regime. To be specific, it modeled the value function using the observed outcome and get a loss function that can be seen as a treatment misclassification loss weighted by the outcome. The OTR is found by minimizing the loss function.

Model:
Model the value function over the observed outcomes:

$$V[d(\mathbf{X})] = E\left[\frac{I(T = d(\mathbf{X}))}{P(T = d(\mathbf{X}) \mid \mathbf{X})}Y\right]$$

Using a simplified version of d(X):$I(z(\mathbf{X}) > 0)$
(the right hand can be seen as misclassification loss when fitting a binary classifier to the actural treatment assignment T)

$$d_{opt}(\mathbf{X}) = \underset{d}{\text{argmin}} E\left[\frac{I(T \neq d(\mathbf{X}))}{P(t \mid \mathbf{X})}Y\right]$$

$$z_{opt}(\mathbf{X}) = \underset{z}{\text{argmin}} E\left[\frac{I(T \neq I(z(\mathbf{X}) > 0))}{P(t \mid \mathbf{X})}Y\right]$$

4.2 ROWSi

Summary: The regularized outcome weighted subgroup identification (Xu et al., 2015) is OWL-based methods that considered the simplicity and interpretability of the treatment assignment rule as the main goal.

multiplicity control: To make the optimal rule d(x) more clinically interpretable and manageable, the estimated rule is always approximated with a simpler rule or set of rules. In OWL, the problem of finding an OTR is modeled as a class of penalized regression problems for binary outcomes. It fit a weighted logistic regression model with the lasso penalty and (negative) binomial loss to the treatment labels.

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{\text{argmin}} \left\{ n^{-1} \sum_{i=1}^{n} L\left[t_i, z\left(\mathbf{x}_i \mid \boldsymbol{\beta}\right)\right] w_i + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

$$w_i = \begin{array}{c} y_i/\pi, \text{ if } t_i = 1, \\ y_i/(1-\pi), \text{ if } t_i = 0, \end{array}$$

z is modeled as a linear predictor for the probability of treatment selection on the logit scale.

$$z\left(\mathbf{x}_i \mid \boldsymbol{\beta}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

optimal treatment assignment rule: (equal to identify subgroup)

$$\widehat{d}^{*}_{opt}(\mathbf{x}) = I(z(\mathbf{x} \mid \widehat{\boldsymbol{\beta}}) > 0)$$

$$\widehat{S} = \{\mathbf{x} : z(\mathbf{x} \mid \widehat{\boldsymbol{\beta}}) > 0\}$$

measure of treatment effect: Introduced measures that summarize the average treatment effect for patients allocated to the treatment and control arms respectively and that can quantify the performance of a treatment assignment rule.

$$d_{+}(\widehat{\boldsymbol{\beta}}) = E\{E(Y \mid z(\mathbf{X} \mid \widehat{\boldsymbol{\beta}}) > 0, T = 1) - E(Y \mid z(\mathbf{X} \mid \widehat{\boldsymbol{\beta}}) > 0, T = 0)\}$$
$$d_{-}(\widehat{\boldsymbol{\beta}}) = E\{E(Y \mid z(\mathbf{X} \mid \widehat{\boldsymbol{\beta}}) < 0, T = 0) - E(Y \mid z(\mathbf{X} \mid \widehat{\boldsymbol{\beta}}) < 0, T = 1)\}$$

d_{+} was the treatment contrast in the subgroup of patients assigned to the treatment arm, that is, patients with z$>$ 0, with larger values indicating a beneficial effect of the experimental treatment.

d_{-} was the treatment contrast in the control arm, that is, z $<$ 0, with larger values indicating a beneficial effect of the control treatment.

bias control: Bootstrap is employed to get the debiased estimation of treatment contrast. This approach helps reduce the overoptimism bias associated with the standard resubstitution estimates.

$$\tilde{d}_{+}\left(\widehat{\boldsymbol{\beta}}^{(b)}\right) = \frac{1}{\left|S_1^{(b)}\right|} \sum_{i \in S_1^{(b)}} y_i - \frac{1}{\left|S_0^{(b)}\right|} \sum_{i \in S_0^{(b)}} y_i$$
$$\tilde{d}_{-}\left(\widehat{\boldsymbol{\beta}}^{(b)}\right) = \frac{1}{\left|\bar{S}_1^{(b)}\right|} \sum_{i \in S_1^{(b)}} y_i - \frac{1}{\left|\bar{S}_0^{(b)}\right|} \sum_{i \in \bar{S}_0^{(b)}} y_i$$

$$S_0^{(b)} = \left\{i : z\left(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{(b)}\right) > 0, t_i = 0\right\}$$
$$S_1^{(b)} = \left\{i : z\left(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{(b)}\right) > 0, t_i = 1\right\}$$
$$\bar{S}_0^{(b)} = \left\{i : z\left(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{(b)}\right) \leqslant 0, t_i = 0\right\}$$
$$\bar{S}_1^{(b)} = \left\{i : z\left(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{(b)}\right) \leqslant 0, t_i = 1\right\}$$

5 Local modeling

Within this approach, the interest lies in studying specific subsets of the space, and there is no longer a need to estimate the outcome function over the entire covariate space.

5.1 PRIM

Summary: Patient rule induction method(PRIM) (Chen et al., 2015) shift the focus from model the outcome to the problem of bump hunting, that is, examining local features of the covariate space, known as bumps, such as regions with a strong treatment effect.
model:

$$\mathrm{EY} = \beta_0 + \beta_1 Z + \beta_2 \mathrm{ZI}(S)$$

5.2 SIDES

Summary: SIDES (Lipkovich et al., 2011) focused on deciding subsets of the space, without estimating the outcome function. And it used the p-value of HTE as a score to split subgroups, and the final subgroup was found through recursive partitioning and confirmed through a certain criterion of the HTE p-value. The subgroup that this method select is the most promising subgroup that has large positive treatment effects.

Model:
splitting criterion c for the i^{th} covariate is selected by maximizing the test statistic $D(c)$ that repersent the significance of treatment effect heterogeneity:

$$D(c) = 2\left[1 - \Phi\left(\frac{|T_H(c) - T_L(c)|}{\sqrt{2}}\right)\right]$$

$$c_i^* = \underset{c \in C_i}{\operatorname{argmin}} D\left(X_i, c\right)$$

$$d_i = D\left(X_i, c_i^*\right)$$

subgroup selection: only retain m promising subgroup based on the value of adjusted splitting criterion. Promising subgroup: the subgroup with larger positive treatment effect.

$$S_i = L_i^* \quad \text{if} \quad T\left(L_j^*\right) > T\left(H_j^*\right)$$

complexity criterion: A promising subgroup is explored further only if the treatment effect in this subgroup is appreciably large compared with the effect in the parent group.

$$p_i \leqslant \gamma p_0$$

multiplicity adjustment: A multiplicity-adjusted p-value for the subgroup Sj is defined as the proportion of null datasets where the treatment difference in the best subgroup is more significant than the treatment difference within Sj.

$$\tilde{p}_j = \frac{1}{K}\sum_{k=1}^{K} I\left\{q_k \leqslant p_j^*\right\}$$

A non-significant multiplicity-adjusted p-value suggested that the apparent treatment effect in the top subgroup was most likely due to selection bias.

5.3 SIDEScreen

Summary: SIDESbase perform best in a relatively small number of candidate biomarkers. The SIDEScreen (Lipkovich and Dmitrienko, 2014) is designed to efficiently handle much larger sets of candidate biomarkers. It's a 2 stage procedure that adds a stage of variable selection using VI score before applying SIDESbase algorithm.

Model:
Stage 1: (variable selection stage) Apply the SIDESbase algorithm at the first stage without complexity control to generate a large collection of promising subgroups. A biomarker screen is introduced at the end of the first stage to filter out the biomarkers that are poor predictors of treatment response using the VI score as a criteria.

Stage 2: the SIDESbase algorithm is applied to the selected biomarkers with stronger predictive properties to arrive at the final set of patient subgroups.

VI score: the average value of the splitting criterion on biomarker i over all subgroups included in the final set.

$$\mathrm{VI}\left(X_i\right) = \frac{1}{m}\sum_{j=1}^{m}\lambda_{ij}, i = 1, \ldots, p$$

Note: VI can quantify a biomarker's average predictive ability

multiplicity control: Adaptive biomarker screen (data-driven threshold derived from the null distribution of the maximum VI score) is employed in stage 1 to select predictive variables. It effectively shrinks the search space and reduces the multiplicity burden and, subsequently, results in a more efficient multiplicity adjustment.

$$\mathrm{VI}(X) \geqslant \hat{E}_0\left(\mathrm{VI}_{\max}\right) + c\sqrt{\hat{V}_0\left(\mathrm{VI}_{\max}\right)}$$

Note: similar to that of min+1SE in tree.

References

Chen, G., Zhong, H., Belousov, A., and Devanarayan, V. (2015). A prim approach to predictive-signature development for patient stratification. *Statistics in Medicine*, 34(2):317–342.

Dusseldorp, E. and Van Mechelen, I. (2014). Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. *Statistics in Medicine*, page 219–237.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, page 2867–2880.

Fu, H., Zhou, J., and Faries, D. E. (2016). Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies. *Statistics in Medicine*, page 3285–3302.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*.

Lipkovich, I. and Dmitrienko, A. (2014). Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using sides. *Journal of Biopharmaceutical Statistics*, page 130–153.

Lipkovich, I., Dmitrienko, A., and B. D'Agostino Sr., R. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, 36(1):136–196.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search–a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, page 2601–2621.

Loh, W., Cao, L., and Zhou, P. (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. *WIREs Data Mining and Knowledge Discovery*, 9(5).

Loh, W.-Y., He, X., and Man, M. (2014). A regression tree approach to identifying subgroups with differential treatment effects. *arXiv: Methodology,arXiv: Methodology*.

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

Shen, J. and He, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association*, page 303–312.

Su, X., Zhou, T., Yan, X., Fan, J., and Yang, S. (2008). Interaction trees with censored survival data. *The International Journal of Biostatistics*, 4(1).

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Xu, Y., Yu, M., Zhao, Y., Li, Q., Wang, S., and Shao, J. (2015). Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics*, page 645–653.