Treatment Recommendation and Parameter Estimation under Single-Index Contrast Function

Menggang Yu

Department of Biostatistics and Medical Informatics
University of Wisconsin - Madison

Singapore, July 2017

Background

- Increasing interest in discovering individualized treatment rules for patients who have different responses to treatments.
  - treatments may be no better than control overall, but may be better for a subgroup of patients with certain characteristics.
- Essentially, we need to investigate interactions between the treatments and covariates to identify the subgroup.

Notations and Assumptions

For the ith patient,

- $T_i = 1$ or 0: treatment indicator
- $Z_i = (1, Z_1, \cdots, Z_p)$: $(p+1)$-vector of predictor variables, including an intercept
- Y_i: observed outcome
- $\pi(Z_i) \equiv P(T_i = 1|Z_i)$: treatment assignment mechanism
- $\pi(T_i|Z_i) = T_i P(T_i = 1|Z_i) + (1 - T_i) P(T_i = 0|Z_i)$

Goal

- Construct a personalized scoring system $f(Z)$
- $f(Z)$ ranks the patients according to the potential treatment effect (or contrast function)

$$\Delta(Z) = \mathbb{E}(Y \mid T = 1, Z) - \mathbb{E}(Y \mid T = 0, Z)$$

- Treatment is recommended for

$$\Omega = \{Z \mid f(Z) > 0\} \approx \{Z \mid \Delta(Z) > 0\}$$

A Simple Interaction Model

- $\mathbb{E}(Y \mid Z, T) = \phi(Z) + T \times f(Z)$
- $\Delta(Z) = \mathbb{E}(Y \mid T = 1, Z) - \mathbb{E}(Y \mid T = 0, Z) = f(Z)$
- It is important to identify the treatment and covariate interactions.
- Main effects of covariates $\phi(Z)$ is in some sense 'separated' from $\Delta(Z)$.

Single Index Model for the Contrast Function

Consider the following single index model

$$\triangle(Z) = \mathbb{E}(Y|T=1,Z) - \mathbb{E}(Y|T=0,Z) = g(\beta^\top Z),$$

- Observe that

$$\mathbb{E}\left[\frac{(2T-1)Y}{\pi(T|Z)}\,\middle|\, Z\right] = \triangle(Z)$$

- Song et al (2007) considered estimation based on

$$\sum_i \left\{\frac{(2T_i-1)Y_i}{\pi(T_i|Z_i)} - g(Z_i\beta)\right\}^2$$

- Assuming $\|\beta\| = 1$ and g from expansion of B-spline basis.

Our Goal

- When g is monotone: even without estimating g, $\beta^\top Z$ is still interpretable in the sense of treatment assignment.
- If we assume g is monotone and the goal is treatment assignment, do we need to estimate g? Will the results be better?
- Is there a systematic way to come up with estimation equations?
- Deal with high dimensional data
- Deal with multiple treatments

Another Estimating Equation

Based on the work of Shuai et al (2017), define the risk

$$R_g(b) = \mathbb{E}\left[\frac{\{Y - (T - 1/2)g(b^\top Z)\}^2}{\pi(T|Z)}\right].$$

- $R_g(b)$ is the expectation of

$$W_Z(b) = \mathbb{E}\left[\left\{Y - 2^{-1}g(b^\top Z)\right\}^2 | T = 1, Z\right] + \mathbb{E}\left[\left\{Y + 2^{-1}g(b^\top Z)\right\}^2 | T = 0, Z\right]$$

- The minimizer of $R_g(b)$ is unique and equal to β if g is second order differentiable and g' is always positive.

Another Estimating Equation

The empirical version of $R_g(b)$ is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\{Y_i-(T_i-1/2)g(b^\top Z_i)\}^2}{\pi(T_i|Z_i)}$$

If g were known, then we could estimate β by finding the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\{Y_i-(T_i-1/2)g(b^\top Z_i)\}}{\pi(T_i|Z_i)}(1-2T_i)g'(b^\top Z_i)Z_i=0,$$

Kernel Weighted Estimating Equation

Note that $g'(0)(b^\top Z)$ is a good approximation to $g(b^\top Z)$ near 0,

$$g(b^\top Z) \approx g(0) + g'(0)(b^\top Z) = g'(0)(b^\top Z).$$

We can estimate $g'(0)\beta$, via the following kernel based version,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\{Y_i - (T_i - 1/2)(b^\top Z_i)\}}{\pi(T_i|Z_i)}(1 - 2T_i)Z_i K_h(b^\top Z_i) = 0.$$

Theoretical Results

- Assume that the kernel K satisfies the usual conditions;
- $h \to 0$ and $nh \to \infty$ as $n \to \infty$;
- Z has a density f;
- $\beta_j \neq 0$ for at least one $j \geq 1$ and without loss of generality $\beta_p \neq 0$.

Theoretical Results

Let $\tilde{b}$ be a solution to the kernel weighted equation. Then, as $n \to \infty$,

(i) $\tilde{b}$ converges in probability to $g'(0)\beta$.

(ii) If $nh^5 \to 0$, then $(nh)^{1/2}\{\tilde{b} - g'(0)\beta\}$ converges in distribution to the p-dimensional normal distribution with mean 0 and covariance matrix Σ.

(iii) The optimal choice of h is $h \asymp n^{-1/5}$, where $a \asymp b$ means $a = O(b)$ and $b = O(a)$.

Dealing with High Dimensional Covariates

When the dimension of Z is very high, we propose to add a LASSO penalty and solve

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\{Y_i-(T_i-1/2)(b^\top Z_i)\}}{\pi(T_i|Z_i)}(1-2T_i)Z_iK_h(b^\top Z_i)+\lambda s(b)=0$$

where

- $\lambda \geq 0$ is a tuning parameter,
- $s(b)$ is the subgradient of $p(b)=\sum_{j=1}^{p}|b_j|$ whose jth component is $\mathrm{sign}(b_j)$ if $b_j \neq 0$ and c if $b_j = 0$, $0 < c < 1$.

Let $\hat{b}$ be a solution to the above equation. We can show that $\hat{b}$ possesses a weak oracle property (Fan and Lv, 2011).

Simulations

We consider respectively the low and high dimensional covariate settings under the following model,

$$Y = (\beta^\top Z/2)^2 + (T - 1/2)g(\beta^\top Z) + \epsilon,$$

where $\epsilon \sim N(0, 0.3^2)$, ϵ, Z and T are independent, and g has the following three forms:

- linear model: $g(\beta^\top Z) = 7\beta^\top Z$
- logistic model: $g(\beta^\top Z) = 7\left\{\exp(\beta^\top Z)/\{1 + \exp(\beta^\top Z)\} - 1/2\right\}$
- probit model: $g(\beta^\top Z) = 7\left\{\Phi(\beta^\top Z) - 1/2\right\}$, where Φ is the standard normal distribution

Simulation - Low Dimensional Setting

- The treatment T takes 0 and 1 with equal probability.
- $p = 3$, $\beta = (1, 1, 1, 1)^\top$, Z_1, Z_2, and Z_3 are independently distributed as the standard normal.
- $n = 200$, 500, and 1000
- Bootstrap variance estimators with bootstrap size 1000.
- All methods produce negligible biases based on 1000 simulation runs.

Root MSE Results in Low Dimensional Case

	Linear			Probit			Logistic		
Estimate	Ours	MCM[†]	FindIt[‡]	Ours	MCM	FindIt	Ours	MCM	FindIt
$n = 200$									
$\tilde{b}_1/\tilde{b}_0$	.013	.041	.071	.109	.350	.132	.126	.418	.134
$\tilde{b}_2/\tilde{b}_0$	.014	.041	.072	.109	.340	.132	.131	.462	.143
$\tilde{b}_3/\tilde{b}_0$	.014	.042	.070	.104	.325	.133	.130	.430	.135
$n = 500$									
$\tilde{b}_1/\tilde{b}_0$	.008	.027	.044	.062	.179	.081	.071	.195	.082
$\tilde{b}_2/\tilde{b}_0$	.008	.027	.046	.059	.164	.078	.075	.206	.084
$\tilde{b}_3/\tilde{b}_0$	.008	.026	.046	.059	.165	.080	.074	.192	.081

[†] Modified Covariate Method (MCM) by Tian et al. (2014)

[‡] FindIt by Imai and Ratkovic (2013).

Coverage Results in Low Dimensional Case

	Linear			Probit			Logistic		
Estimate	Ours	MCM	FindIt	Ours	MCM	FindIt	Ours	MCM	FindIt
$n = 200$									
$\tilde{b}_1/\tilde{b}_0$	.947	.949	.939	.943	.942	.959	.950	.947	.951
$\tilde{b}_2/\tilde{b}_0$	.948	.944	.942	.945	.951	.951	.960	.940	.961
$\tilde{b}_3/\tilde{b}_0$	.953	.951	.939	.941	.956	.957	.943	.951	.956
$n = 500$									
$\tilde{b}_1/\tilde{b}_0$	.953	.923	.960	.950	.947	.952	.955	.943	.938
$\tilde{b}_2/\tilde{b}_0$	.948	.923	.938	.955	.948	.953	.945	.955	.947
$\tilde{b}_3/\tilde{b}_0$	.947	.955	.955	.952	.945	.938	.945	.945	.957

Simulation - High Dimensional Setting

- $\beta = (\beta_0, \cdots, \beta_p)^\top$, where $p = 23$, $\beta_j = 1$, $j = 0, 1, 2, 3$, and $\beta_j = 0$ for $j \geq 4$
- $Z_1, ..., Z_p$ are still independently distributed as the standard normal
- Other setting are the same as that for the low dimensional case.

Coverage Results in High Dimensional Case

	Linear			Probit			Logistic		
Estimate	Ours	MCM	FindIt	Ours	MCM	FindIt	Ours	MCM	FindIt
$n = 200$									
$\hat{b}_1/\hat{b}_0$	0.958	0.948	0.944	0.958	0.868	0.962	0.946	0.920	0.948
$\hat{b}_2/\hat{b}_0$	0.948	0.939	0.949	0.962	0.895	0.962	0.967	0.930	0.949
$\hat{b}_3/\hat{b}_0$	0.953	0.932	0.948	0.954	0.923	0.890	0.960	0.925	0.793
$n = 500$									
$\hat{b}_1/\hat{b}_0$	0.944	0.977	0.954	0.949	0.925	0.937	0.947	0.948	0.966
$\hat{b}_2/\hat{b}_0$	0.944	0.943	0.971	0.941	0.948	0.966	0.954	0.977	0.937
$\hat{b}_3/\hat{b}_0$	0.947	0.937	0.948	0.960	0.954	0.977	0.962	0.954	0.971

Variable Selection Results in High Dimensional Case

	Linear			Probit			Logistic		
n	Ours	MCM	FindIt	Ours	MCM	FindIt	Ours	MCM	FindIt
200	0.988	0.984	0.691	0.913	0.804	0.295	0.828	0.706	0.378
500	1.000	1.000	0.897	1.000	0.962	0.814	0.997	0.972	0.894

A Flexible Semiparametric Model

The single index contrast function model is equivalent to

$$Y = \frac{1}{2} T g(\beta^\top Z) + \epsilon,$$

- g is an unknown function,
- ϵ satisfies

$$E\left[\frac{T}{\pi(T|Z)} \epsilon(Z) \middle| Z \right] = 0.$$

This model is equivalent to the following model

$$Y = h(T, \beta^\top Z) + \epsilon(Z),$$

where $\epsilon(Z)$ is some random variable satisfying the above equation.

Semiparametric Efficiency Theory

The likelihood of (Z, T, Y) is

$$\eta_{10}(Z) \times \pi(T|Z) \times \eta_{20}\{Y - \frac{1}{2}Tg(\beta_0^\top Z), Z, T\}$$

- $\eta_{10}(\cdot)$ is the density of Z
- $\eta_{20}(\cdot)$ is the density of ϵ conditional on Z and T
- Note that η_{10}, η_{20}, and g are infinite-dimensional nuisance parameters.
- $\pi(T|Z)$ is either known or estimated by a parametric model

Semiparametric Efficiency Theory

The orthogonal complement of the nuissance tangent space is $\mathcal{S}_0 = \mathcal{S}_{10} \oplus \mathcal{S}_2$,

- $$\mathcal{S}_{10} = \left\{ W_T \{\alpha(Z) - \mathbb{E}[\alpha(Z)|\beta_0^\top Z]\} \left[\epsilon - \frac{\mathbb{E}(W_T^2 \epsilon|Z)}{\mathbb{E}(W_T^2|Z)}\right] : \forall \alpha(Z) \right\},$$

- $$\mathcal{S}_2 = \left\{ W_T \gamma(Z) : \forall \gamma(Z) \right\}.$$

where

$$W_T \equiv \frac{T}{\pi(T|Z)}$$

Semiparametric Efficiency Theory

For any function $h(\epsilon, Z, T)$, its projection on $\mathcal{S}_0$ is given by

$$W_T\, C(Z) \left\{ \epsilon - \frac{\mathbb{E}[W_T^2 \epsilon | Z]}{\mathbb{E}[W_T^2 | Z]} \right\} + W_T \frac{\mathbb{E}[W_T h | Z]}{\mathbb{E}[W_T^2 | Z]},$$

where

- $$C(Z) = W_{Z,T,\epsilon} \left\{ D(Z) - \frac{\mathbb{E}[W_{Z,T,\epsilon} D(Z) | \beta_0^\top Z]}{\mathbb{E}[W_{Z,T,\epsilon} | \beta_0^\top Z]} \right\},$$
- $$W_{Z,T,\epsilon} = \left\{ \mathbb{E}[W_T^2 \epsilon^2 | Z] - \frac{\mathbb{E}[W_T^2 \epsilon | Z]^2}{\mathbb{E}[W_T^2 | Z]} \right\}^{-1},$$
- $$D(Z) = \mathbb{E}[W_T h \epsilon | Z] - \frac{\mathbb{E}[W_T^2 \epsilon | Z] \mathbb{E}[W_T h | Z]}{\mathbb{E}[W_T^2 | Z]}.$$

Efficient score

The efficient score is

$$W_T C(Z)\left\{\epsilon - \frac{\mathbb{E}[W_T^2\epsilon|Z]}{\mathbb{E}[W_T^2|Z]}\right\},$$

where

$$C(Z) = W_{Z,T,\epsilon}\, g^{'}(\beta_0^\top Z)\left\{Z - \frac{\mathbb{E}[W_{Z,T,\epsilon} Z|\beta_0^\top Z]}{\mathbb{E}[W_{Z,T,\epsilon}|\beta_0^\top Z]}\right\},$$

and,

$$W_{Z,T,\epsilon} = \left\{\mathbb{E}[W_T^2\epsilon^2|Z] - \frac{\mathbb{E}[W_T^2\epsilon|Z]^2}{\mathbb{E}[W_T^2|Z]}\right\}^{-1}.$$

Efficiency Considerations

- In general, the efficient score is very hard to estimate directly
- Choose the most efficient estimating equation in a smaller subspace in the nuisance tangent space, e.g.,

$$\tilde{S} = \left\{ W_T g'(\beta_0^\top Z) Z^\top \{\epsilon - \eta(Z)\}, \forall \eta(Z) \right\}.$$

- Choices for $\eta(Z)$
  - $\eta(Z) = 0$, adopted in our estimation.
  - $\eta(Z) = \{1 - 2\pi(Z)\} g(\beta_0^\top Z)$ for Song et al. (2017).
  - $\eta(Z) = \frac{E[W_T^2 \epsilon | Z]}{E[W_T^2 | Z]}$ leads to the most efficient estimator in $\tilde{S}$ for any function g.

Efficiency Considerations

When $\eta(Z)$ does not need to be estimated, we propose minimizing the following loss function

$$\frac{\{Y - \frac{1}{2}Tg(\beta_0^\top Z) - \eta(Z)\}^2}{\pi(T|Z)}.$$

Formally, our MAVE-type estimator with η, tMAVE$_\eta$, is defined to be the minimizer of the following optimization function:

$$\begin{aligned} & L(\beta, \{a_j, b_j\}_{j=1}^n) \\ =& \frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n \frac{\{Y_i - \frac{1}{2}T_i[a_j + b_j(\beta^\top Z_i - \beta^\top Z_j) - \eta(Z_i)]\}^2}{\pi(T_i|Z_i)} w_{ij}, \end{aligned}$$

where $w_{ij} = K_h(\beta^\top Z_j - \beta^\top Z_i)$.

Efficiency Considerations

A two step estimation process

- Firstly, we solve a tMAVE_0 (with $\eta = 0$), then g and the residuals, $\hat{\epsilon}_i$'s, are estimated by kernel method.
- Then, we solve tMAVE_{eff} by estimating $\frac{E[W_T^2 \epsilon | Z]}{E[W_T^2 | Z]}$ by

$$\frac{\hat{E}[W_T^2 \hat{\epsilon} | Z]}{E[W_T^2 | Z]} = \pi(Z)(1 - \pi(Z)) \frac{\sum_{i=1}^n {K^e}_{h_e}(Z_i - Z) W_{T_i}^2 \hat{\epsilon}_i}{\sum_{i=1}^n {K^e}_{h_e}(Z_i - Z)} \quad (1)$$

where

- $K_{h_e}^e$ is a kernel function with $K_{h_e}^e(Z) = h_e^{-p} K^e(Z / h_e^p)$.
- K^e can be different from the kernel used in tMAVE.

Simulations

$$y = (\beta^\top Z)^2 + (T - 1/2)g(\beta^\top Z) + \epsilon,$$

- Almost the same as previous simulation setting;
- Main effect quadrupled;
- SD of the error term doubled: $\sigma = 0.6$.

Coefficient Estimation with $n = 200$

	Linear		Gaussian		Logistic	
	tMAVE_0	tMAVE_{eff}	tMAVE_0	tMAVE_{eff}	tMAVE_0	tMAVE_{eff}
mean						
$\hat{\beta}_2/\hat{\beta}_1$	0.9995	0.9986	0.8630	0.9161	0.7797	0.8611
$\hat{\beta}_3/\hat{\beta}_1$	1.0021	1.0021	0.8960	0.9410	0.8192	0.8884
$\hat{\beta}_4/\hat{\beta}_1$	1.0042	1.0035	0.8891	0.9408	0.8013	0.8802
$\sqrt{mse}$						
$\hat{\beta}_2/\hat{\beta}_1$	0.0563	0.0378	0.3122	0.2044	0.4106	0.2890
$\hat{\beta}_3/\hat{\beta}_1$	0.0586	0.0386	0.2971	0.1977	0.4056	0.2837
$\hat{\beta}_4/\hat{\beta}_1$	0.0540	0.0361	0.3075	0.2055	0.4191	0.2847

Coefficient Estimation with $n = 500$

	Linear		Gaussian		Logistic	
	tMAVE$_0$	tMAVE$_{eff}$	tMAVE$_0$	tMAVE$_{eff}$	tMAVE$_0$	tMAVE$_{eff}$
mean						
$\hat{\beta}_2/\hat{\beta}_1$	0.9978	0.9994	0.9526	0.9759	0.8995	0.9484
$\hat{\beta}_3/\hat{\beta}_1$	1.0010	1.0004	0.9701	0.9854	0.9193	0.9625
$\hat{\beta}_4/\hat{\beta}_1$	1.0020	1.0004	0.9452	0.9798	0.8994	0.9477
$\sqrt{mse}$						
$\hat{\beta}_2/\hat{\beta}_1$	0.0372	0.0207	0.1676	0.0975	0.2539	0.1558
$\hat{\beta}_3/\hat{\beta}_1$	0.0329	0.0188	0.1663	0.0935	0.2587	0.1507
$\hat{\beta}_4/\hat{\beta}_1$	0.0326	0.0184	0.1675	0.0925	0.2531	0.1505

Rank Correlation with $n = 200$

	Linear	Gaussian	Logistic
Single	0.9893(0.0121)	0.6318(0.3266)	0.5435(0.3582)
tMAVE_0	0.9893(0.0122)	0.6675(0.2897)	0.5919(0.3191)
tMAVE_0(index)	0.9983(0.0018)	0.8707(0.2224)	0.8255(0.2734)
tMAVE_{eff}	0.9903(0.0122)	0.6887(0.3072)	0.6086(0.3450)
tMAVE_{eff}(index)	0.9993(0.0008)	0.9406(0.1564)	0.9077(0.1952)
$W^{\dagger}_{sq-L}$	0.9909(0.0093)	0.6319(0.4688)	0.5608(0.5183)
$W^{\dagger}_{sq-A}$	0.9608(0.0284)	0.5722(0.2709)	0.5079(0.3035)
$W^{\dagger}_{flo-L}$	0.9823(0.0249)	0.4348(0.3457)	0.3760(0.3649)

† These methods are based on Chen et al (2017).

Rank Correlation with $n = 500$

	Linear	Gaussian	Logistic
Single	0.9935(0.0071)	0.8013(0.2159)	0.7377(0.2657)
tMAVE_0	0.9937(0.0072)	0.7941(0.1978)	0.7342(0.2436)
tMAVE_0(index)	0.9994(0.0006)	0.9503(0.1375)	0.9088(0.2004)
tMAVE_{eff}	0.9941(0.0072)	0.8129(0.1879)	0.7496(0.2704)
tMAVE_{eff}(index)	0.9998(0.0002)	0.9931(0.0245)	0.9800(0.0612)
W_{sq-L}	0.9965(0.0029)	0.8244(0.2721)	0.7612(0.3486)
W_{sq-A}	0.9807(0.0126)	0.7108(0.1774)	0.6523(0.2107)
W_{flo-L}	0.9935(0.0072)	0.5972(0.2641)	0.5311(0.2945)

Correct Classification Rate with $n = 200$

	Linear	Gaussian	Logistic
Single	0.9771(0.0087)	0.7810(0.1600)	0.7291(0.1662)
tMAVE_0	0.9788(0.0085)	0.8177(0.1442)	0.7725(0.1543)
tMAVE_0(index)	0.9841(0.0080)	0.8781(0.1147)	0.8502(0.1302)
tMAVE_{eff}	0.9832(0.0072)	0.8347(0.1554)	0.7847(0.1697)
tMAVE_{eff}(index)	0.9897(0.0052)	0.9242(0.0838)	0.8983(0.0999)
W_{sq-L}	0.9559(0.0186)	0.7393(0.1806)	0.7107(0.1947)
W_{sq-A}	0.7915(0.0332)	0.5347(0.0391)	0.5271(0.0353)
W_{flo-L}	0.9417(0.0331)	0.6370(0.1178)	0.6173(0.1196)

Correct Classification Rate with $n = 500$

	Linear	Gaussian	Logistic
Single	0.9861(0.0052)	0.8857(0.1101)	0.8398(0.1367)
tMAVE_0	0.9876(0.0049)	0.8972(0.1006)	0.8567(0.1222)
tMAVE_0(index)	0.9905(0.0045)	0.9347(0.0787)	0.9026(0.1031)
tMAVE_{eff}	0.9909(0.0040)	0.9157(0.1086)	0.8738(0.1415)
tMAVE_{eff}(index)	0.9946(0.0026)	0.9713(0.0223)	0.9527(0.0413)
W_{sq-L}	0.9721(0.0108)	0.8232(0.1191)	0.7927(0.1425)
W_{sq-A}	0.7880(0.0207)	0.5193(0.0227)	0.5145(0.0198)
W_{flo-L}	0.9629(0.0165)	0.6927(0.0989)	0.6672(0.1033)

Conclusion and Discussion

- It is natural to extend the single index model to multiple index model

$$\Delta(Z) := \mathbb{E}[Y|T=1,Z] - \mathbb{E}[Y|T=0,Z] = g(B_0^\top Z),$$

  where B_0 is a $p \times d$ matrix.

- Extension to multiple treatment is not trivial:
  - Choice of g: g_1, $g_2, \ldots$
  - Choice of β and model consistency: single vs. multiple index?
  - Treatment ordered?

Acknowledgement

Joint work with

- Cui Xiong and Jun Shao
- Muxuan Liang