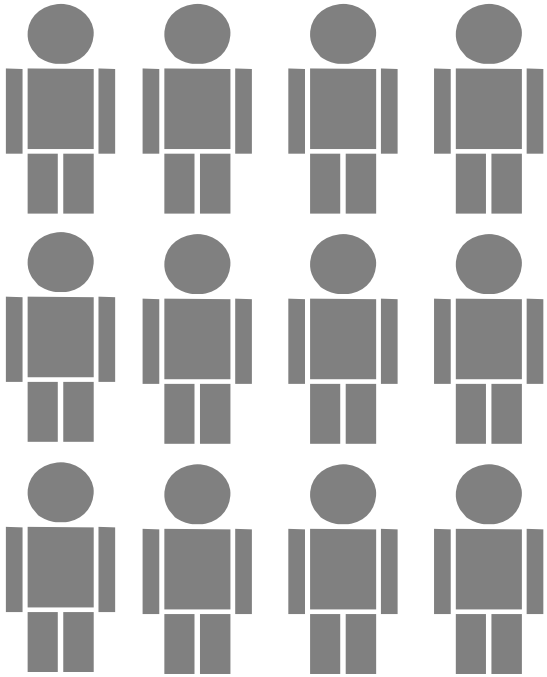
A Sparse Random Projection-based Test for Overall Qualitative Treatment Effects

Chengchun Shi

Department of Statistics
North Carolina State University

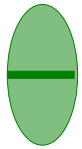
Joint work with Wenbin Lu and Rui Song, JASA, accepted

One-size-fits-all



Patients

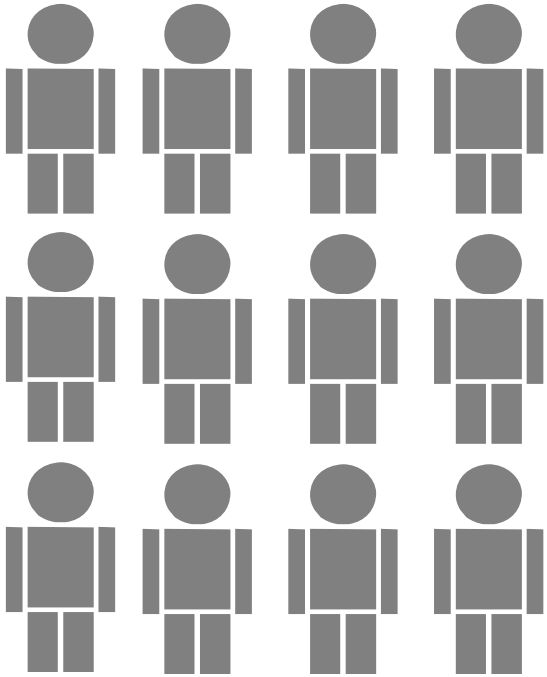
Treatment A



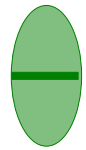
Treatment B

Treatment C

One-size-fits-all

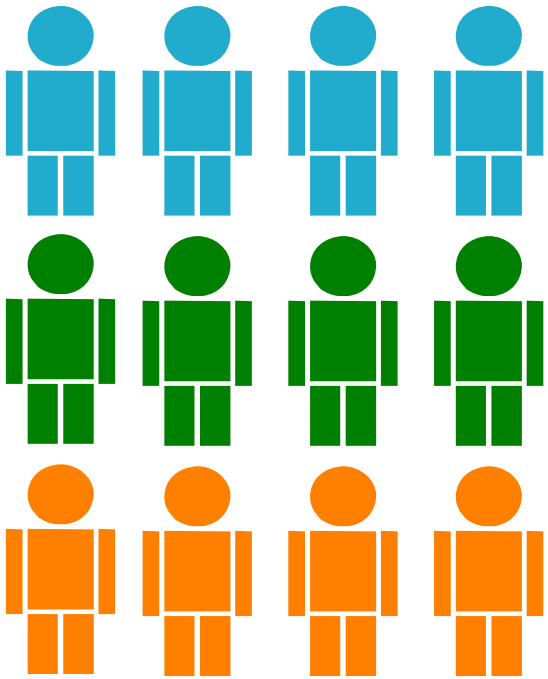


Patients

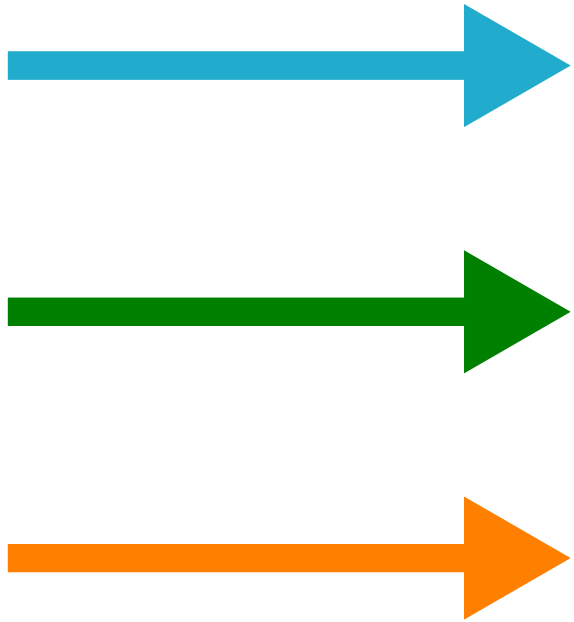


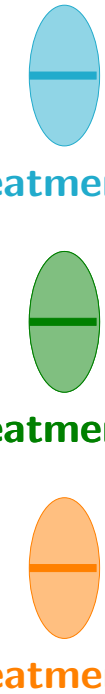
Treatment B

Precision medicine: individualized treatment regime (ITR)



Patients





Treatment A

Treatment B

Treatment C

Estimation of the optimal ITR

- Q-learning (Watkins and Dayan, 1992; Murphy, 2005; Moodie et al., 2014; Song et al., 2015)
- A-learning (Murphy, 2003; Robins, 2004; Schulte et al., 2014)
- Value search method (Zhang et al., 2012, 2013)
- Outcome weighted learning (OWL, Zhao et al., 2012, 2015)
- Decision lists (Zhang et al., 2015, 2017)
- Tree-based methods (Laber and Zhao, 2015; Zhu et al., 2017)

Inference of the optimal ITR

- All the above estimation methods implicitly assume patients' covariates have qualitative interactions with the treatment
- Testing the overall qualitative treatment effects (OQTE):
  H_0: No OQTE $\Longleftrightarrow$ Implementing the optimal ITR is the same as "one-size-fits-all".

Testing the overall qualitative treatment effects (OQTE)

- Help determine whether to "individualize" or not. If H_0 is not rejected:
  - The optimal ITR may not be of practical interest;
  - Recommend "one-size-fits-all" in settings patients' covariates are expensive to obtain (Baker et al., 2009; Gail, 2009; Huang et al., 2015).
- Offer guidance on constructing the optimal ITR. If H_0 is rejected:
  - Identify covariates with qualitative treatment effects.

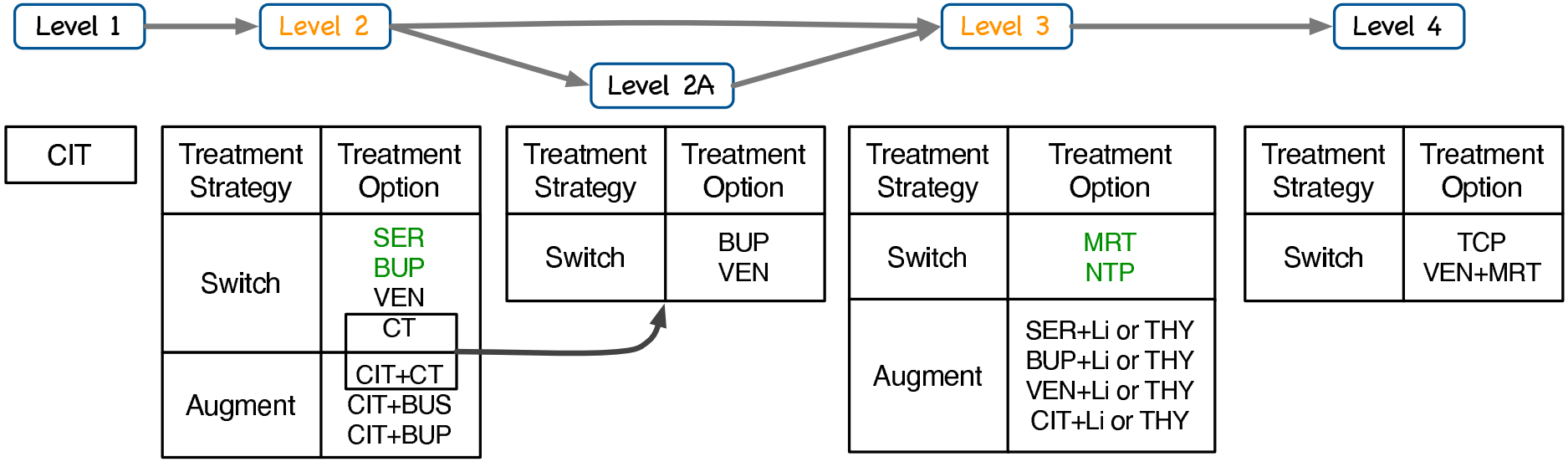
Exisiting literature

- Medical studies: Gail and Simon (1985); Roth and Simon (2018).
- Education: Chang et al. (2015); Hsu (2017).
- All focus on fixed p scenario

Nefazodone-CBASP clinical trial study

- 681 patients with nonpsychotic chronic major depressive disorder
- **Treatments:**
  - Nefazodone;
  - Cognitive Behavioral-Analysis System of Psychotherapy (CBASP);
  - the combination of Nefazodone and CBASP.
- **Response:** the negative 24-item Hamilton Rating Scale for Depression (HRSD).
- **Covariates:** 647 patients with complete records of 50 covariates.

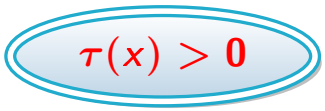
Sequenced Treatment Alternatives to Relieve Depressions (STAR*D) Study



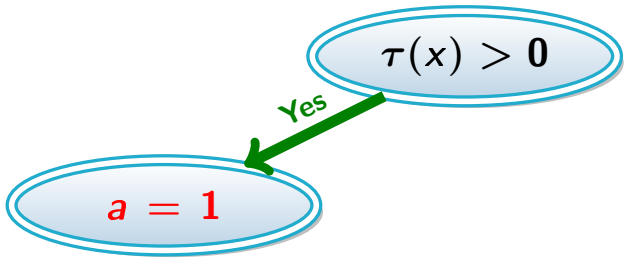
- 381 covariates available at Level 3, 305 at Level 2
- 73 patients BUP or SER at Level 2, MRT or NTP at Level 3

- **Individualized treatment regime (ITR)**: a function maps the covariate space to the space of available treatment options.
- **Optimal ITR**: yields the most favorable clinical outcome.
- **Single stage study**:
  - **Covariates** X: p-dimensional vector.
  - **Treatment** A: 1 for the treatment, 0 for the control.
  - **Response** Y: assuming larger values are desirable.
- Q**-function**: $Q(a, x) = \mathrm{E}(Y|A = a, X = x)$.
- **Contrast function**: $\tau(x) = Q(1, x) - Q(0, x)$.

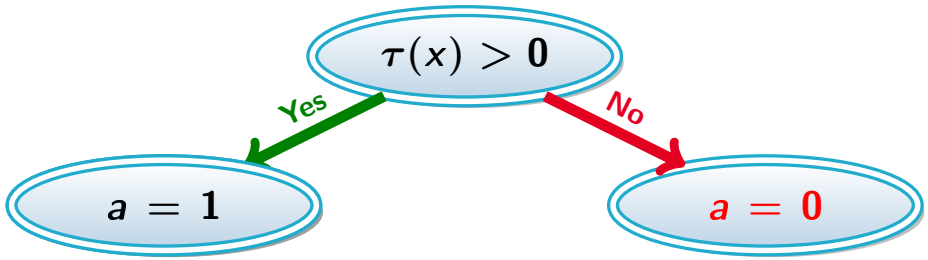
- **Individualized treatment regime (ITR)**: a function maps the covariate space to the space of available treatment options.
- **Optimal ITR**: yields the most favorable clinical outcome.
- **Single stage study**:
  - **Covariates** X: p-dimensional vector.
  - **Treatment** A: 1 for the treatment, 0 for the control.
  - **Response** Y: assuming larger values are desirable.
- Q**-function**: $Q(a, x) = \mathrm{E}(Y|A = a, X = x)$.
- **Contrast function**: $\tau(x) = Q(1, x) - Q(0, x)$.



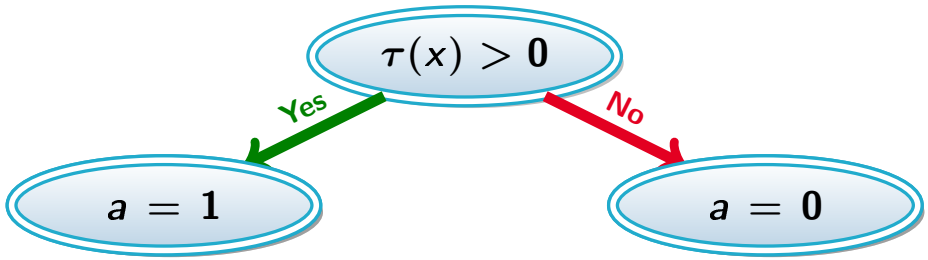
- **Individualized treatment regime (ITR)**: a function maps the covariate space to the space of available treatment options.
- **Optimal ITR**: yields the most favorable clinical outcome.
- **Single stage study**:
  - **Covariates** X: p-dimensional vector.
  - **Treatment** A: 1 for the treatment, 0 for the control.
  - **Response** Y: assuming larger values are desirable.
- Q**-function**: $Q(a, x) = \mathrm{E}(Y|A = a, X = x)$.
- **Contrast function**: $\tau(x) = Q(1, x) - Q(0, x)$.



- **Individualized treatment regime (ITR)**: a function maps the covariate space to the space of available treatment options.
- **Optimal ITR**: yields the most favorable clinical outcome.
- **Single stage study**:
  - **Covariates** X: p-dimensional vector.
  - **Treatment** A: 1 for the treatment, 0 for the control.
  - **Response** Y: assuming larger values are desirable.
- Q**-function**: $Q(a, x) = \mathrm{E}(Y|A = a, X = x)$.
- **Contrast function**: $\tau(x) = Q(1, x) - Q(0, x)$.



- **Individualized treatment regime (ITR)**: a function maps the covariate space to the space of available treatment options.
- **Optimal ITR**: yields the most favorable clinical outcome.
- **Single stage study**:
  - **Covariates** X: p-dimensional vector.
  - **Treatment** A: 1 for the treatment, 0 for the control.
  - **Response** Y: assuming larger values are desirable.
- Q**-function**: $Q(a, x) = \mathrm{E}(Y|A = a, X = x)$.
- **Contrast function**: $\tau(x) = Q(1, x) - Q(0, x)$.



- $d^{opt}(x) = \mathbb{I}(\tau(x) > 0)$

- X has **OQTE** iff $\Pr\{\tau(X) > 0\} > 0$ and $\Pr\{\tau(X) < 0\} > 0$.
- $H_{0,b}$: X does not have OQTE v.s $H_{1,b}$: X has OQTE.
- Define the **value difference function**,

$$\mathrm{VD}_a(d) \equiv \mathrm{E}Y^*(d) - \mathrm{E}Y^*(a) = \mathrm{E}\tau(X)\{d(X) - a\},$$

  for a given ITR $d(\cdot) : X \to \{0, 1\}$ and $a \in \{0, 1\}$.
- Suppose $\mathrm{E}|\tau(X)| < +\infty$. Testing $H_{0,b}$ is equivalent to

$$H_{0,c} : \min_{a\in\{0,1\}} \mathrm{VD}_a(d^{opt}) = 0 \text{ v.s } H_{1,c} : \min_{a\in\{0,1\}} \mathrm{VD}_a(d^{opt}) > 0.$$

- Suppose Treatment 1 is on average better than Treatment 0.

$$H_{0,d} : \mathrm{VD}_1(d^{opt}) = 0 \text{ v.s } H_{1,d} : \mathrm{VD}_1(d^{opt}) > 0.$$

- Estimate $\mathrm{VD}_1(d)$ based on (A)IPWE (Zhang et al., 2012)
- Challenge: consistently estimate d^{opt} in high dimensions

- X has **OQTE** iff $\Pr\{\tau(X) > 0\} > 0$ and $\Pr\{\tau(X) < 0\} > 0$.
- **$H_{0,b}$: X does not have OQTE v.s $H_{1,b}$: X has OQTE.**
- Define the **value difference function**,

$$\mathrm{VD}_a(d) \equiv \mathrm{E}Y^*(d) - \mathrm{E}Y^*(a) = \mathrm{E}\tau(X)\{d(X) - a\},$$

  for a given ITR $d(\cdot) : X \to \{0, 1\}$ and $a \in \{0, 1\}$.
- Suppose $\mathrm{E}|\tau(X)| < +\infty$. Testing $H_{0,b}$ is equivalent to

$$H_{0,c} : \min_{a\in\{0,1\}} \mathrm{VD}_a(d^{opt}) = 0 \text{ v.s } H_{1,c} : \min_{a\in\{0,1\}} \mathrm{VD}_a(d^{opt}) > 0.$$

- Suppose Treatment 1 is on average better than Treatment 0.

$$H_{0,d} : \mathrm{VD}_1(d^{opt}) = 0 \text{ v.s } H_{1,d} : \mathrm{VD}_1(d^{opt}) > 0.$$

- Estimate $\mathrm{VD}_1(d)$ based on (A)IPWE (Zhang et al., 2012)
- Challenge: consistently estimate d^{opt} in high dimensions

- X has **OQTE** iff $\Pr\{\tau(X) > 0\} > 0$ and $\Pr\{\tau(X) < 0\} > 0$.
- $H_{0,b}$: X does not have OQTE v.s $H_{1,b}$: X has OQTE.
- Define the **value difference function**,

$$\mathrm{VD}_a(d) \equiv \mathrm{E}Y^*(d) - \mathrm{E}Y^*(a) = \mathrm{E}\tau(X)\{d(X) - a\},$$

  for a given ITR $d(\cdot) : X \to \{0, 1\}$ and $a \in \{0, 1\}$.
- Suppose $\mathrm{E}|\tau(X)| < +\infty$. Testing $H_{0,b}$ is equivalent to

$$\mathbf{H_{0,c}} : \min_{\mathbf{a}\in\{0,1\}} \mathrm{VD}_{\mathbf{a}}(\mathbf{d^{opt}}) = \mathbf{0} \text{ v.s } \mathbf{H_{1,c}} : \min_{\mathbf{a}\in\{0,1\}} \mathrm{VD}_{\mathbf{a}}(\mathbf{d^{opt}}) > \mathbf{0}.$$

- Suppose Treatment 1 is on average better than Treatment 0.

$$H_{0,d} : \mathrm{VD}_1(d^{opt}) = 0 \text{ v.s } H_{1,d} : \mathrm{VD}_1(d^{opt}) > 0.$$

- Estimate $\mathrm{VD}_1(d)$ based on (A)IPWE (Zhang et al., 2012)
- Challenge: consistently estimate d^{opt} in high dimensions

- X has **OQTE** iff $\Pr\{\tau(X) > 0\} > 0$ and $\Pr\{\tau(X) < 0\} > 0$.
- $H_{0,b}$: X does not have OQTE v.s $H_{1,b}$: X has OQTE.
- Define the **value difference function**,

$$\mathrm{VD}_a(d) \equiv \mathrm{E}Y^*(d) - \mathrm{E}Y^*(a) = \mathrm{E}\tau(X)\{d(X) - a\},$$

  for a given ITR $d(\cdot) : X \to \{0, 1\}$ and $a \in \{0, 1\}$.
- Suppose $\mathrm{E}|\tau(X)| < +\infty$. Testing $H_{0,b}$ is equivalent to

$$H_{0,c} : \min_{a\in\{0,1\}} \mathrm{VD}_a(d^{opt}) = 0 \text{ v.s } H_{1,c} : \min_{a\in\{0,1\}} \mathrm{VD}_a(d^{opt}) > 0.$$

- Suppose Treatment 1 is on average better than Treatment 0.

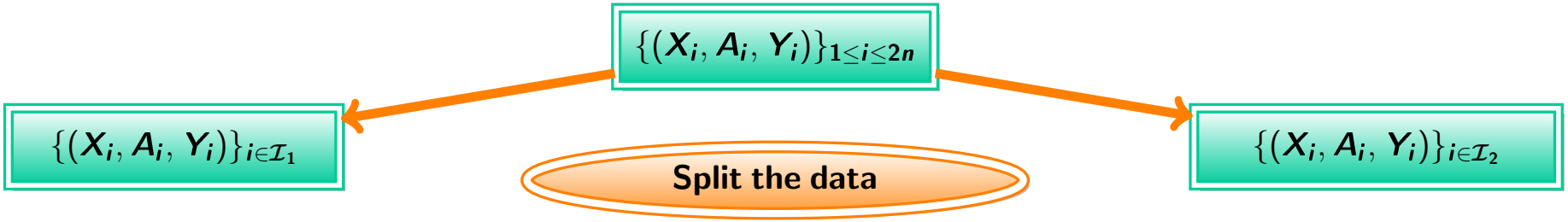
$$\mathbf{H_{0,d}} : \mathrm{VD}_1(\mathbf{d^{opt}}) = \mathbf{0} \text{ v.s } \mathbf{H_{1,d}} : \mathrm{VD}_1(\mathbf{d^{opt}}) > \mathbf{0}.$$

- Estimate $\mathrm{VD}_1(d)$ based on (A)IPWE (Zhang et al., 2012)
- Challenge: consistently estimate d^{opt} in high dimensions

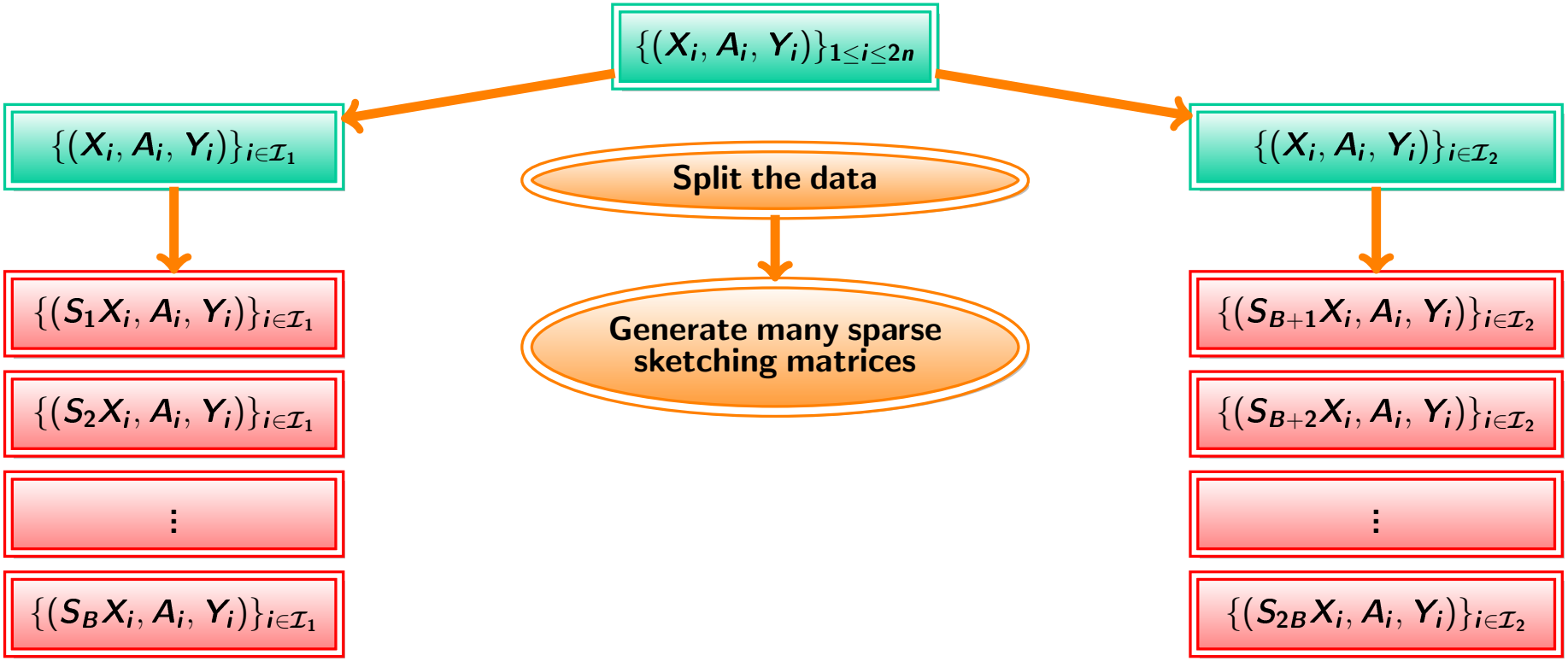
A sparse random projection-based test

$$\{(X_i, A_i, Y_i)\}_{1 \le i \le 2n}$$

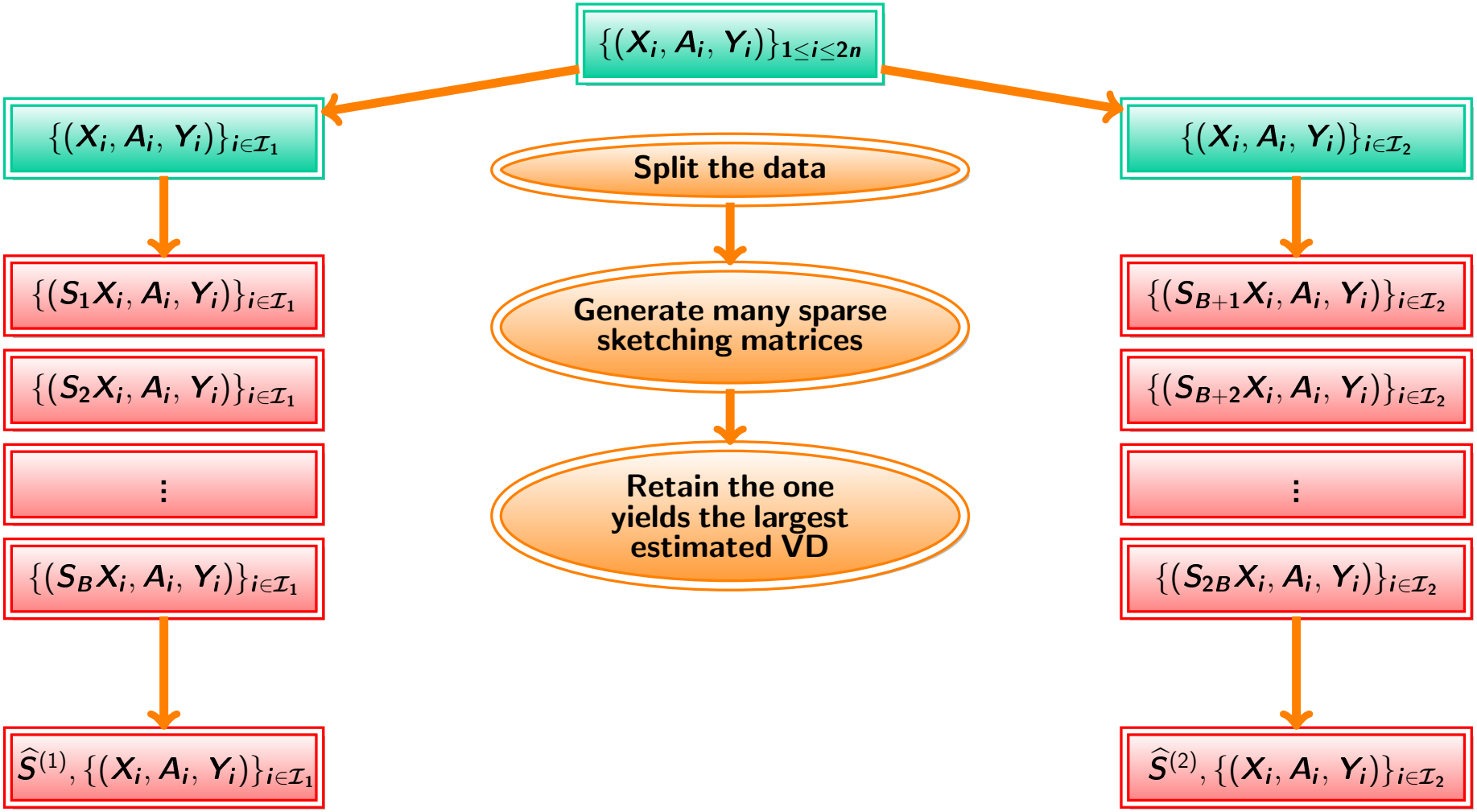
A sparse random projection-based test



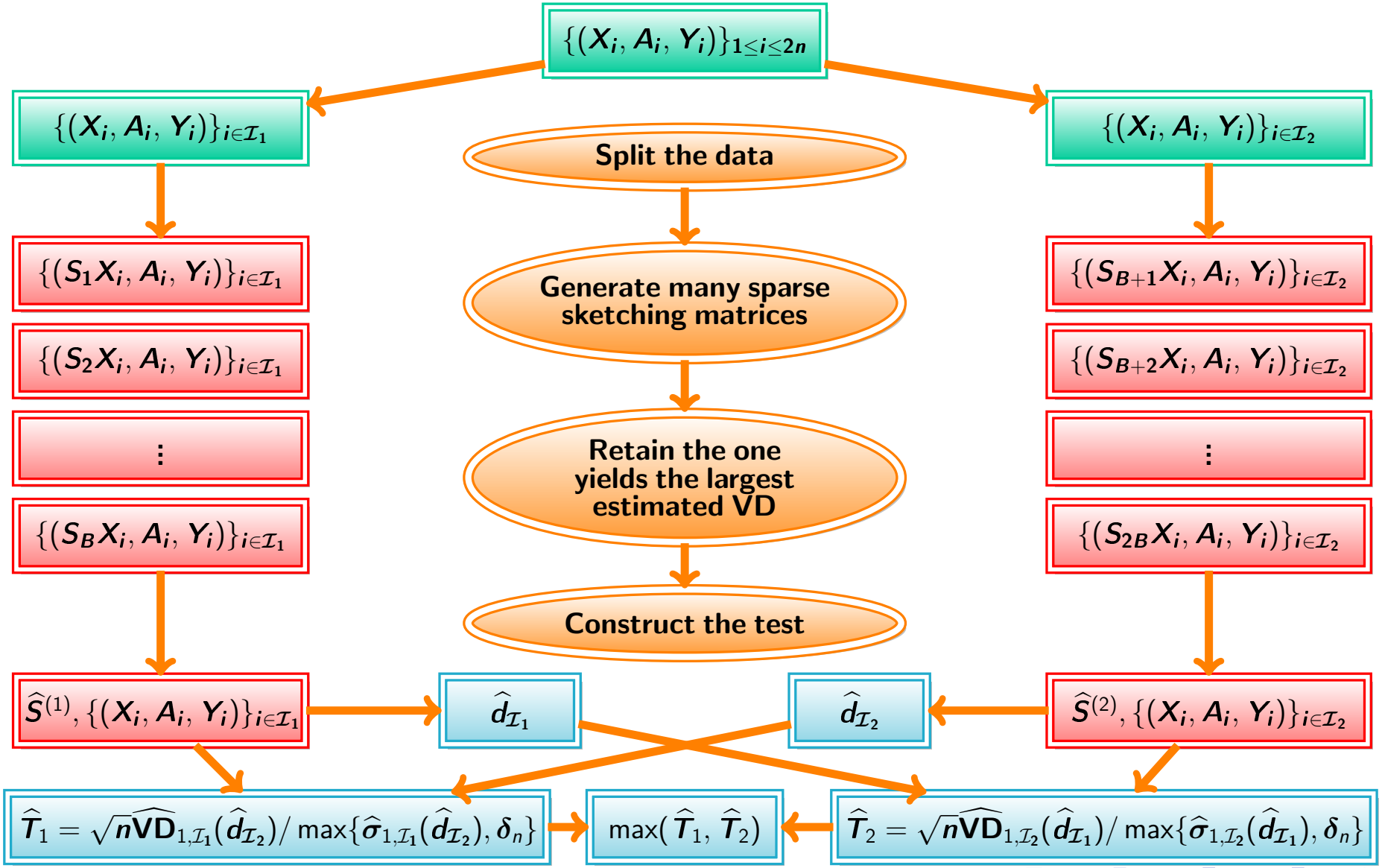
A sparse random projection-based test



A sparse random projection-based test



A sparse random projection-based test



A sparse random projection-based test

- What are sparse sketching matrices?

$$\mathcal{S}(s,q) = \{S \in \mathbb{R}^{q\times p} : \|S^{(j)}\|_0 \le s, \|S^{(j)}\|_2 = 1, \forall j = 1,\ldots,q\}.$$

For $j = 1,\ldots,q$,

  - Select a simple random sample $\mathcal{M}_j$ of size s from $\{1,\ldots,p\}$;
  - Generate a Gaussian random vector $g_j \sim N(0, I_s)$;
  - Set $S^{(j)}_{0,\mathcal{M}_j^c} = 0$ and $S^{(j)}_{0,\mathcal{M}_j} = g_j/\|g_j\|_2$.

- Why sparse sketching matrices?
  - The test statistics are more powerful compared to those based on dense sketching matrices when the optimal ITR is "sparse".

A sparse random projection-based test

- Data-dependent algorithms to generate sparse sketching matrices:
  - Select the matrix that maximizes the cross-validated value difference estimator.
  - Most random matrices are weakly correlated with the contrast function.
  - Improve the power of the test.
- Sample-splitting and cross-validation:
  - Use difference samples to estimate the optimal ITR and evaluate the value difference.
  - Guarantee the validity of the test, even in the nonregular cases.

Under the null hypothesis

$$\widehat{T}_{SRP} = \max\left(\frac{\widehat{\text{VD}}_{1,\mathcal{I}_1}(\widehat{d}_{\mathcal{I}_2})}{\max\{\widehat{\sigma}_{1,\mathcal{I}_1}(\widehat{d}_{\mathcal{I}_2}), \delta_n\}}, \frac{\widehat{\text{VD}}_{1,\mathcal{I}_2}(\widehat{d}_{\mathcal{I}_1})}{\max\{\widehat{\sigma}_{1,\mathcal{I}_2}(\widehat{d}_{\mathcal{I}_1}), \delta_n\}}\right),$$

where

$$\widehat{\text{VD}}_{1,\mathcal{I}}(d) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left(\frac{A_i}{\pi(X_i)} - \frac{1-A_i}{1-\pi(X_i)}\right) Y_i\{1-d(X_i)\},$$

and $\widehat{\sigma}^2_{1,\mathcal{I}}(d)$ denotes its sampling variance estimator.

Theorem

Assume SUTVA, no unmeasured confunders and the positivity assumption holds, $E|Y|^3 = O(1)$, and $\delta_n \gg n^{-1/6}$. Then under H_0,

$$\limsup_n Pr\left(\widehat{T}_{SRP} > z_{\alpha/2}\right) \leq \alpha.$$

Under the alternative hypothesis

- Estimated optimal ITR based on $\{(SX_i, A_i, Y_i)\}_{i\in\mathcal{I}}$
  - The optimal ITR in the projected covariates space:

$$d_S^{opt} = \underset{d:SX\to\{0,1\}}{\arg\max}\ \mathrm{VD}_1(d) = \mathbb{I}\{\tau^S(SX) > 0\},$$

    where $\tau^S(\omega) = \mathrm{E}\{\tau(X)|SX = \omega\}$.
  - Plug-in classifiers:

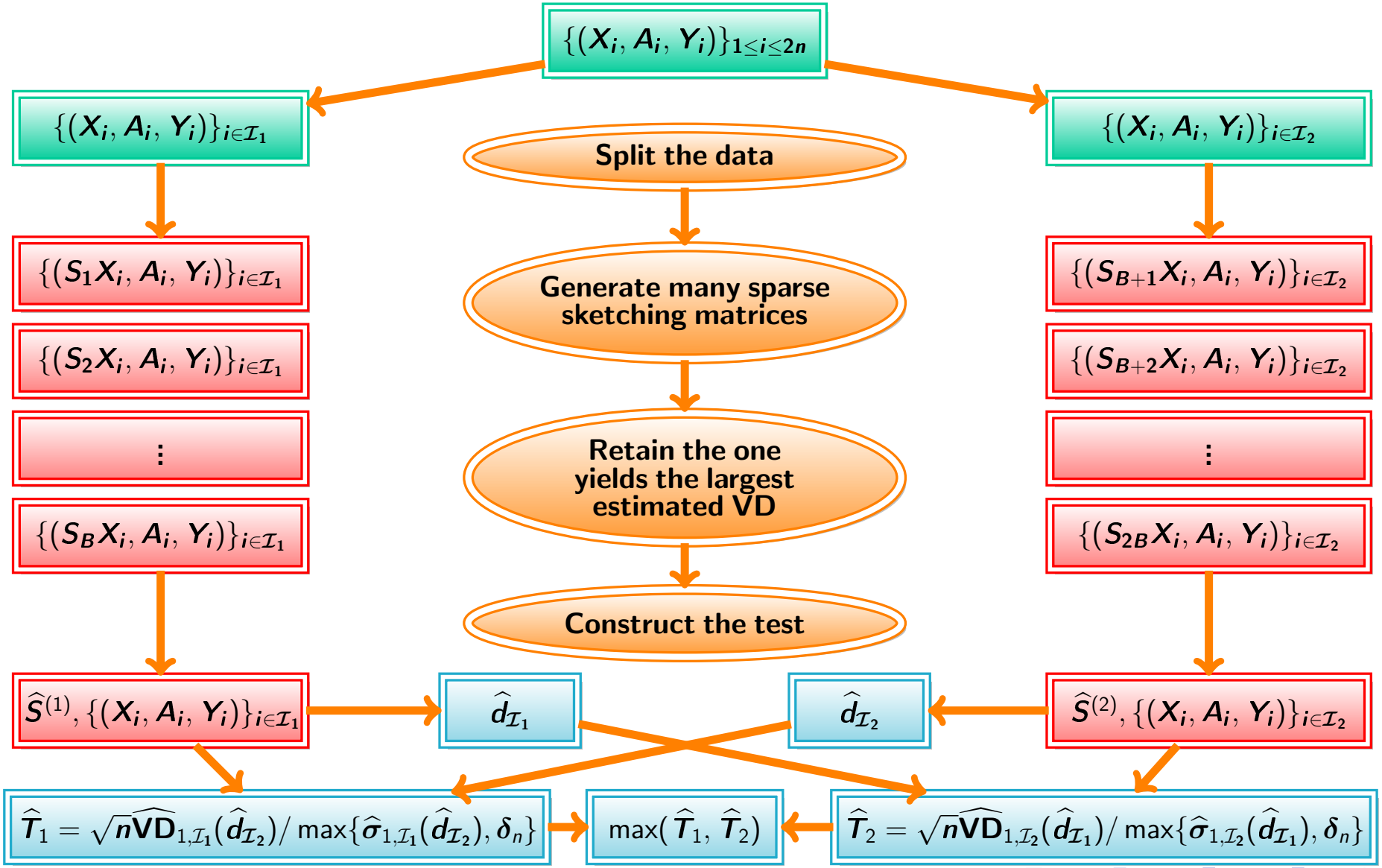
$$\widehat{d}_{\mathcal{I}}^S = \mathbb{I}\{\widehat{\tau}_{\mathcal{I}}^S(SX) > 0\},$$

    $\widehat{\tau}_{\mathcal{I}}^S$ estimated by nonparametric methods such as spline regression.
- Optimal sketching matrix:

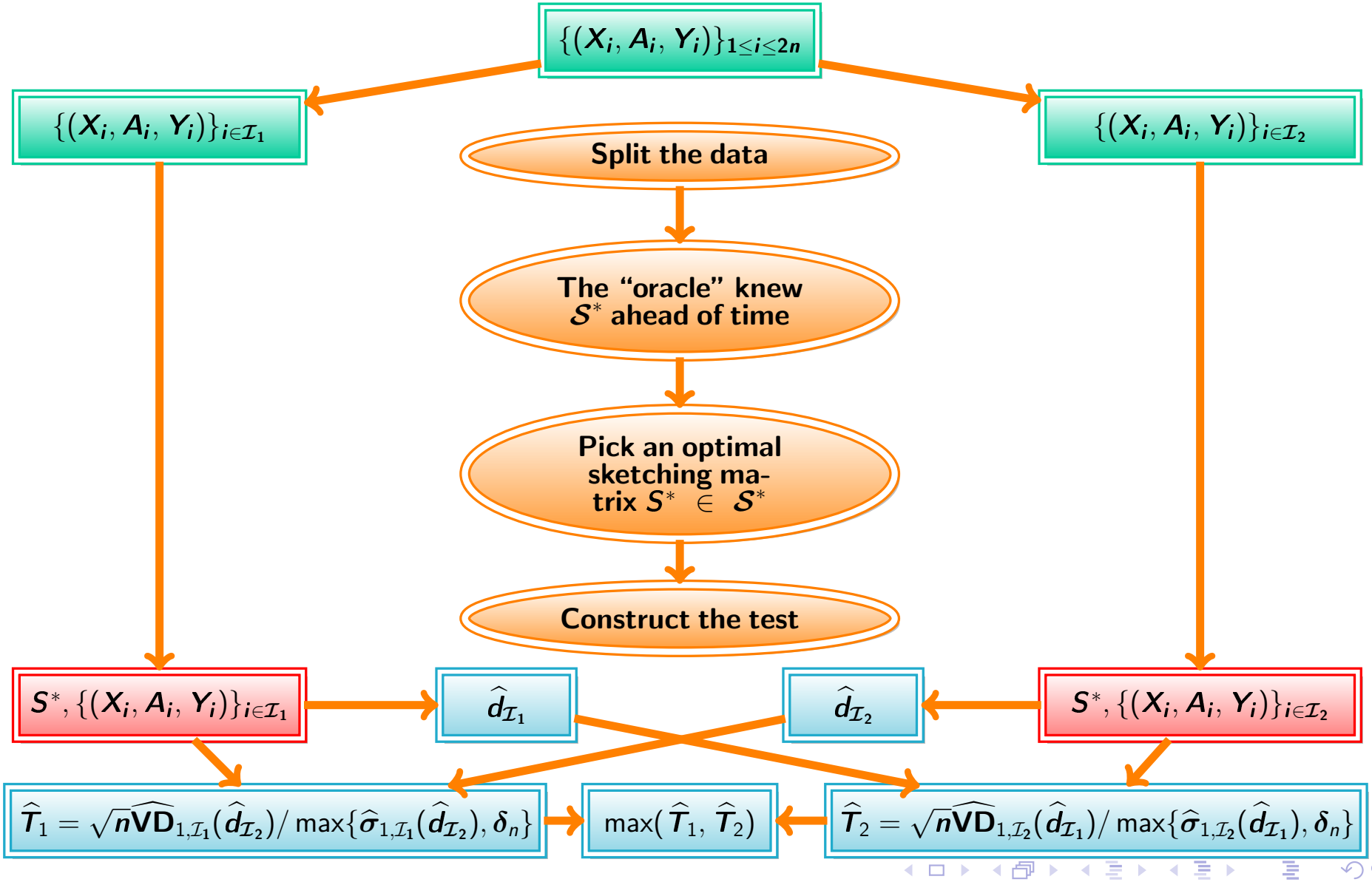
$$S^* \in \underset{S\in\mathcal{S}(s,q)}{\arg\max}\, \mathrm{VD}_1(d_S^{opt}),$$

$\mathcal{S}^*$ denotes the set consisting of all optimal sketching matrices.

The "oracle" test statistic



The "oracle" test statistic



Under the alternative hypothesis

(A1) For any $S_1, S_2, \ldots, S_B \in \mathcal{S}$ and $\mathcal{I} \in \{1, \ldots, 2n\}$, the following occurs with probability tending to 1,

$$\max_{b \in \{1,\ldots,B\}} \mathrm{E}^X |\widehat{\tau}_{\mathcal{I}}^{S_b}(S_b X) - \tau^{S_b}(S_b X)|^2 = O(n^{-r_0} \log n).$$

(A2) $B \gg (p\sqrt{n})^{(s-1)q}$. There exists some $\bar{C} > 0$ and $S^* \in \mathcal{S}^*$ such that

$$\mathrm{E}|\tau^S(SX) - \tau^{S^*}(S^* X)|^2 \leq \bar{C}\|S - S^*\|_F^2.$$

(A3) For any S that satisfies $\mathrm{VD}_1(d_S^{opt}) \geq \mathrm{VD}_1(d) + \varepsilon_0$ for some $\varepsilon_0 > 0$,

$$\Pr\{0 < |\tau^S(SX)| \leq t\} + O(t^\gamma).$$

(A4) $\mathrm{VD}_1(d^{opt}) = \mathrm{VD}_1(d_{S^*}^{opt})$ for some $S^* \in \mathcal{S}^*$.

Under the alternative hypothesis

Theorem ("oracle" property)

Assume conditions in Theorem 1 hold. Assume (A1)-(A2) hold and $\log B = o(n^{1/3})$. *If* $VD_1(d^{opt}_{S^*}) \gg \max(\sqrt{\log B}/\sqrt{n}, n^{-r_0/2}\sqrt{\log n})$, *then*

$$Pr(\widehat{T}_{SRP} > z_{\alpha/2}) \to 1.$$

Further assume (A3), (A4) hold, $Pr\{\tau(X) = 0\} = 0$, $B = O(n^{\kappa_B})$ *for some* $\kappa_B > 0$, $VD_1(d^{opt}_{S^*}) = O(n^{-1/2})$, $r_0 > (\gamma + 2)/(2\gamma + 2)$ *and* $\liminf_n \sigma_0^2 > 0$. *Then,*

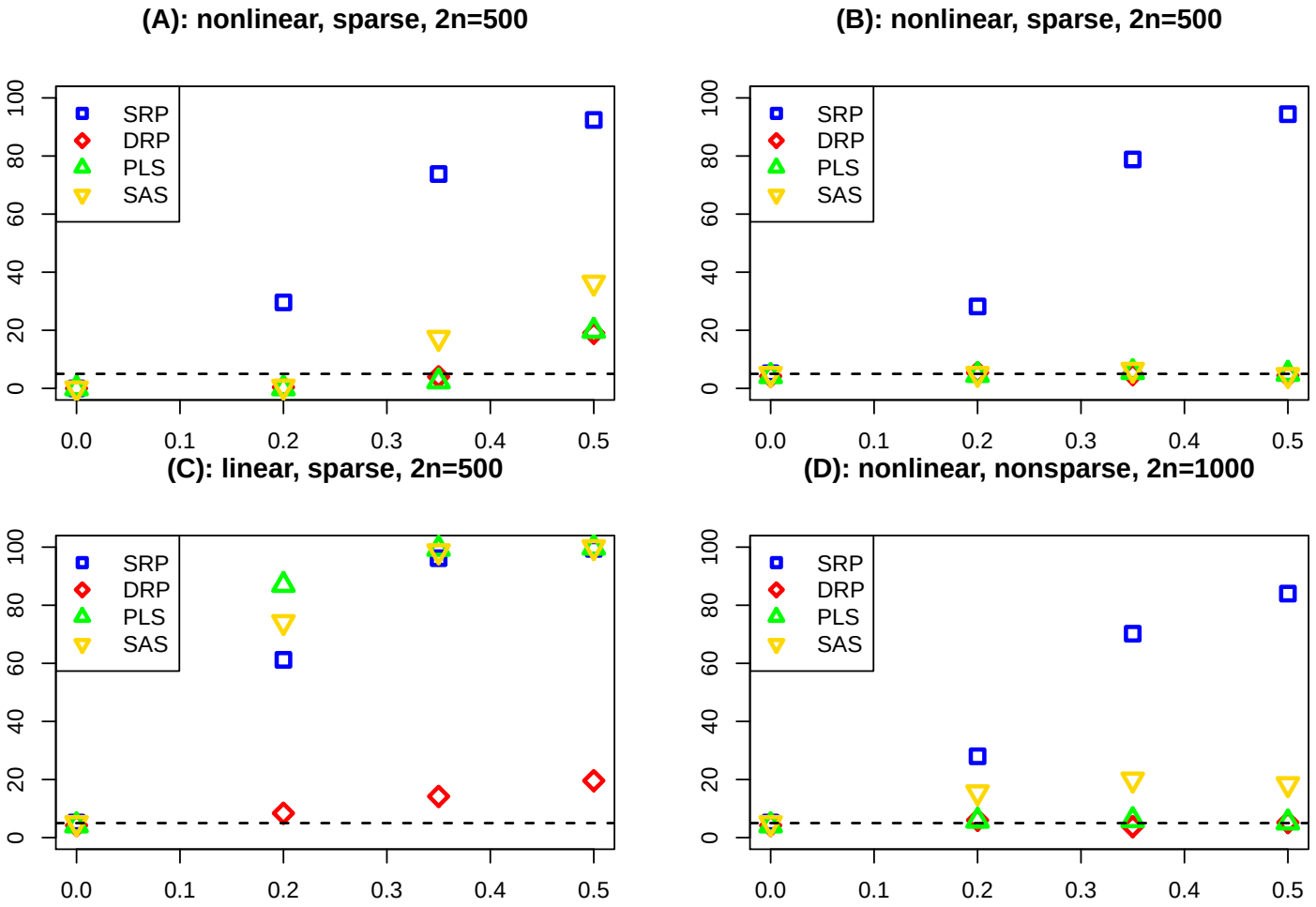
$$Pr(\widehat{T}_{SRP} > z_{\alpha/2}) = Pr(\widehat{T}_{oracle} > z_{\alpha/2}) + o(1),$$

where σ_0^2 *is the asymptotic variance of* $\sqrt{|\mathcal{I}|}\widehat{VD}_1(d^{opt})$.

Some implementation issues

- Doubly-robust value difference estimator:
  - Fit baseline based on penalized linear regression.
  - Fit propensity score based on penalized logistic regression.
- Choice of s and q:
  - Randomly select s and q when generate sketching matrices.
- Choice of B: $B \asymp n^{\kappa_n} p^{\kappa_p}$ for some $\kappa_n, \kappa_p > 0$.

Simulations ($p = 100$, $B = 4 \times 10^6$)



- simulation program written in C with GNU Scientific Library (GSL, Galassi et al., 2015).

Nefazodone-CBASP clinical trial study (Keller et al., 2000)

- 681 patients with nonpsychotic chronic major depressive disorder
- **Treatments:**
  - Nefazodone;
  - Cognitive Behavioral-Analysis System of Psychotherapy (CBASP);
  - the combination of Nefazodone and CBASP.
- **Response:** the negative 24-item Hamilton Rating Scale for Depression (HRSD).
- **Covariates:** 647 patients with complete records of 50 covariates.
- **Pairwise comparison**: Nefazodone v.s the combination; CBASP v.s the combination. Bonferroni's procedure for multiple comparison.
- The null of no OQTE is not rejected. Our tests formally verify findings of Zhao et al. (2012).

Extensions

- Nonnegative average treatment effects.
- Multi-stage studies.
- Conditional qualitative treatment effects.
- Tech companies, online randomized experiment (AB test).

Thank you!