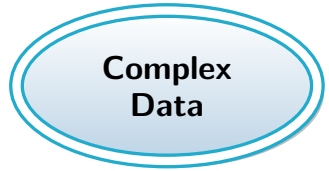
On Statistical Learning for Individualized Decision Making with Complex Data

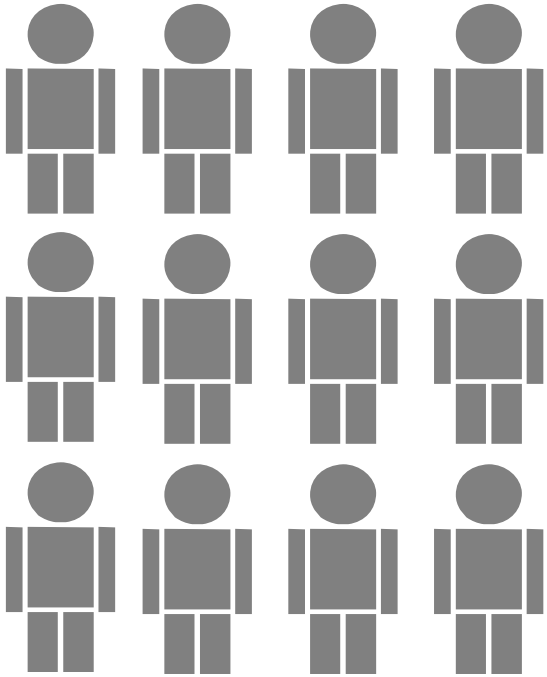
Chengchun Shi

Department of Statistics
North Carolina State University


Individualized
Decision Making

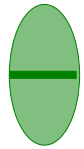

Complex
Data

One-size-fits-all



Patients

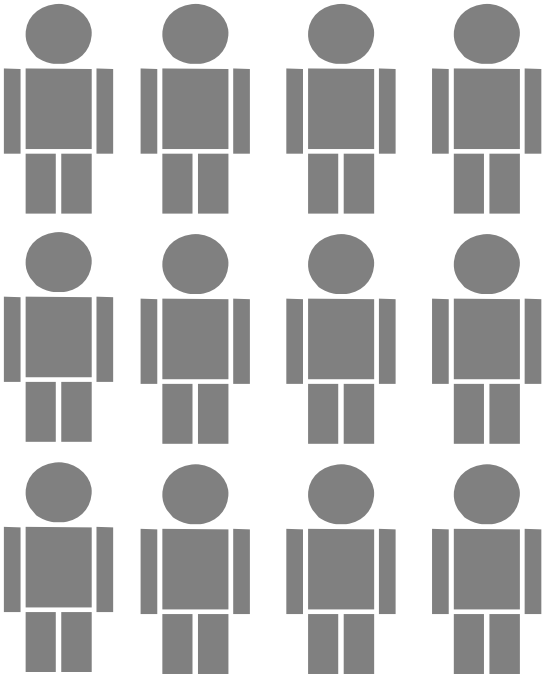
Treatment A



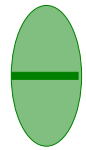
Treatment B

Treatment C

One-size-fits-all

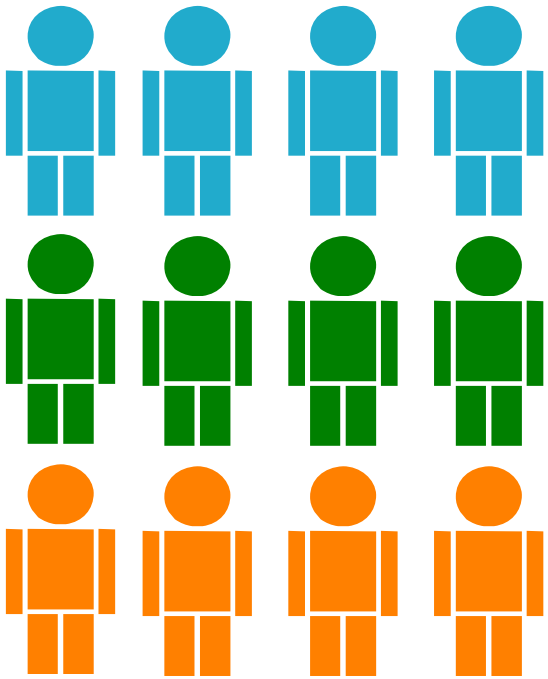


Patients

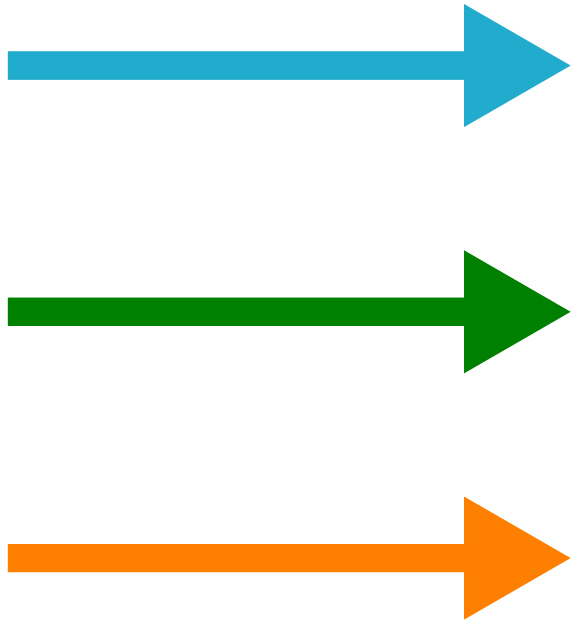


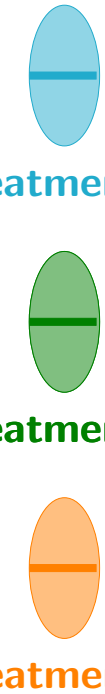
Treatment B

Precision medicine: individualized treatment regime (ITR)



Patients



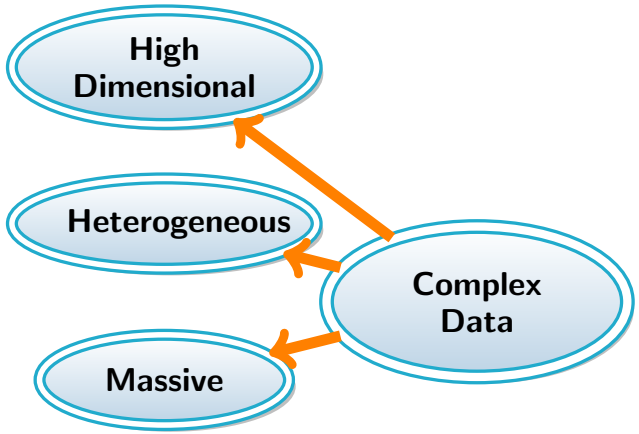


Treatment A

Treatment B

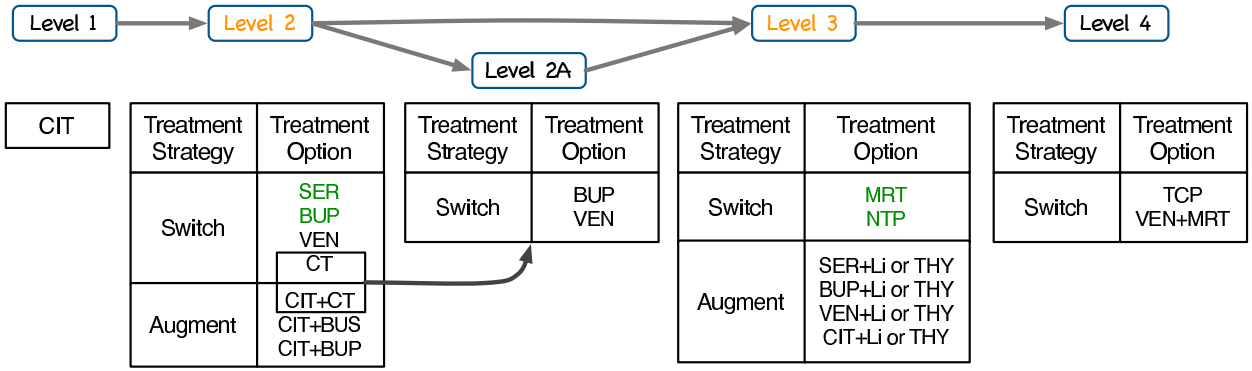
Treatment C

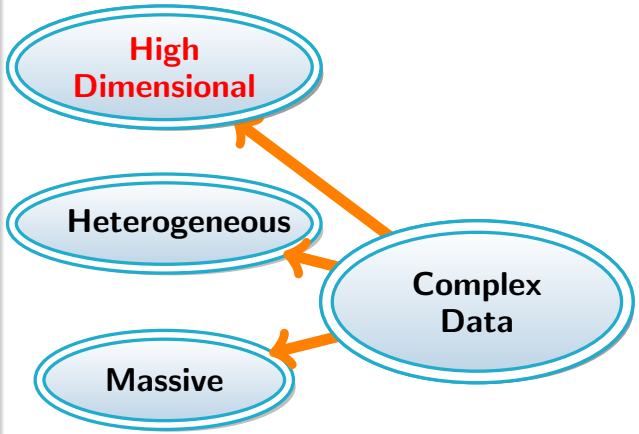

Individualized
Decision Making


High
Dimensional
Heterogeneous
Complex
Data
Massive

Individualized Decision Making

Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Study (Fava et al., 2003)

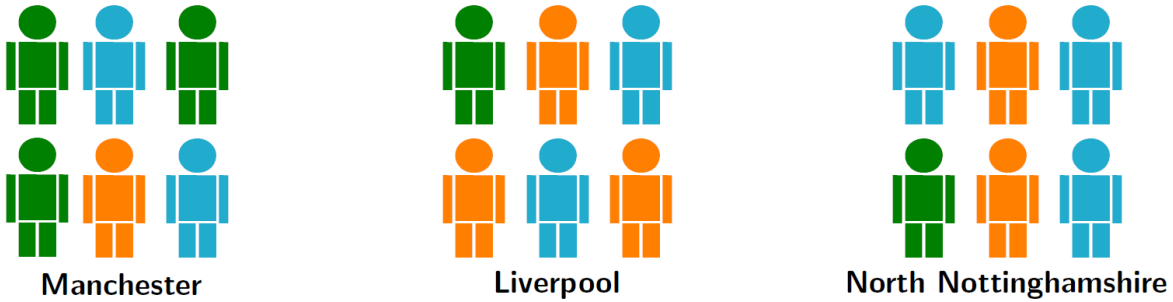




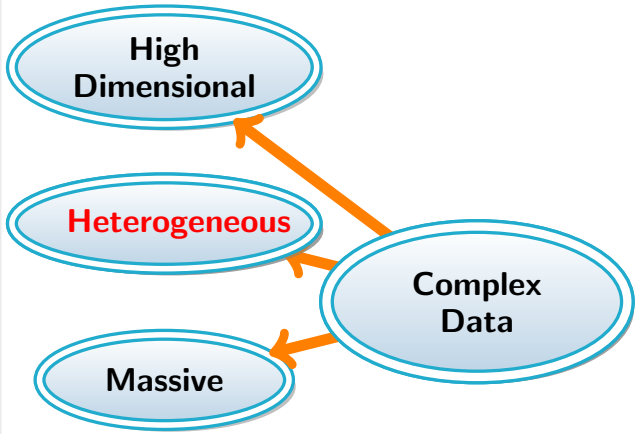
- 381 covariates available at Level 3, 305 at Level 2
- 73 patients BUP or SER at Level 2, MRT or NTP at Level 3

Schizophrenia Study (Tarrier et al., 2004)

- A multicentre, randomized controlled trial
- Over 400 patients with schizophrenia enrolled in 3 treatment centres in England

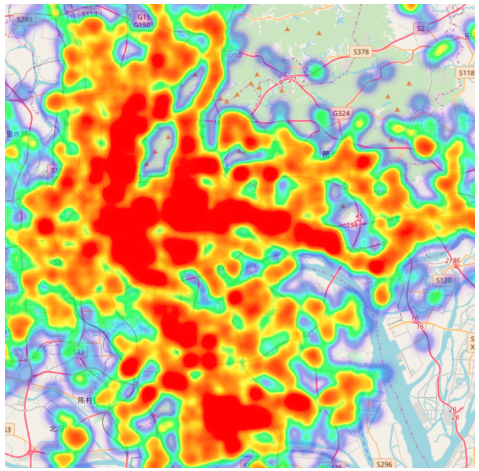


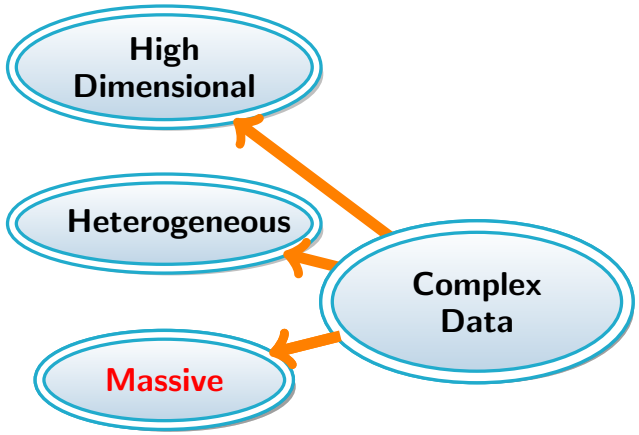
- Treatments: cognitive behavioral therapy (CBT) + treatment as usual; supportive counselling (SC) + treatment as usual.
- Response: the reduction of Positive and Negative Syndrome Scale (PANSS) score after 18 months.
- Covariates: PANSS score at baseline; log duration of untreated psychosis.
- Patients enrolled at different treatment centres show heterogeneity in optimal treatment decision, due to differences in characteristics of treatment setting across centres

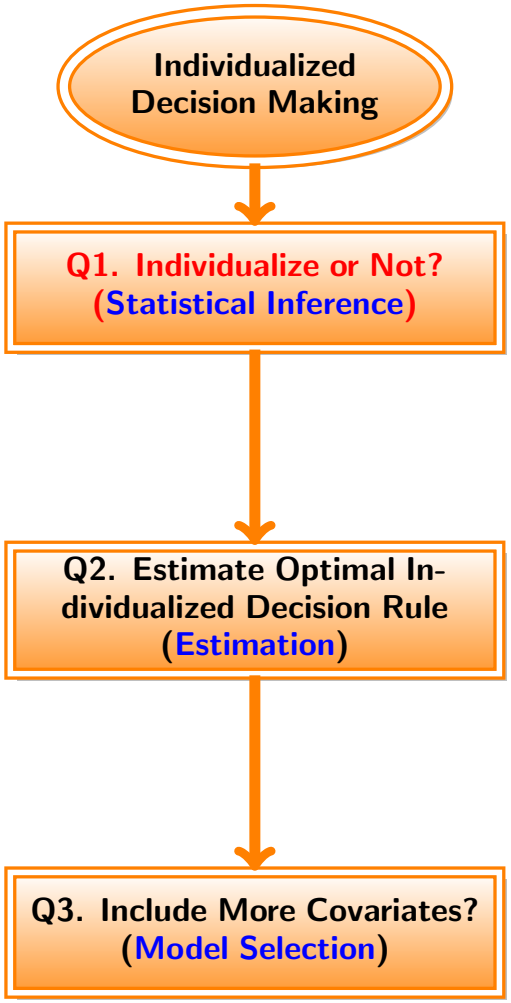


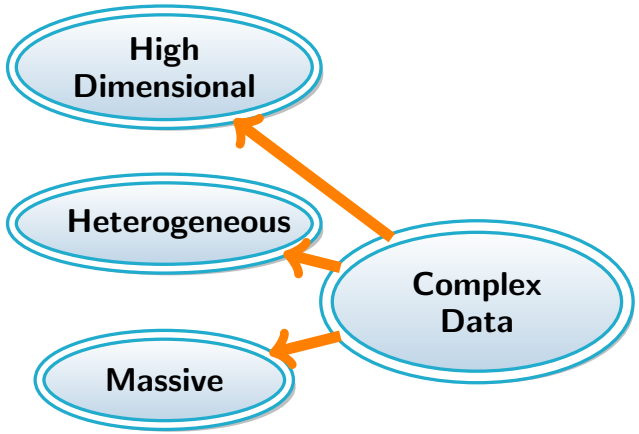
Didi Chuxing (the world's leading ridesharing platform)

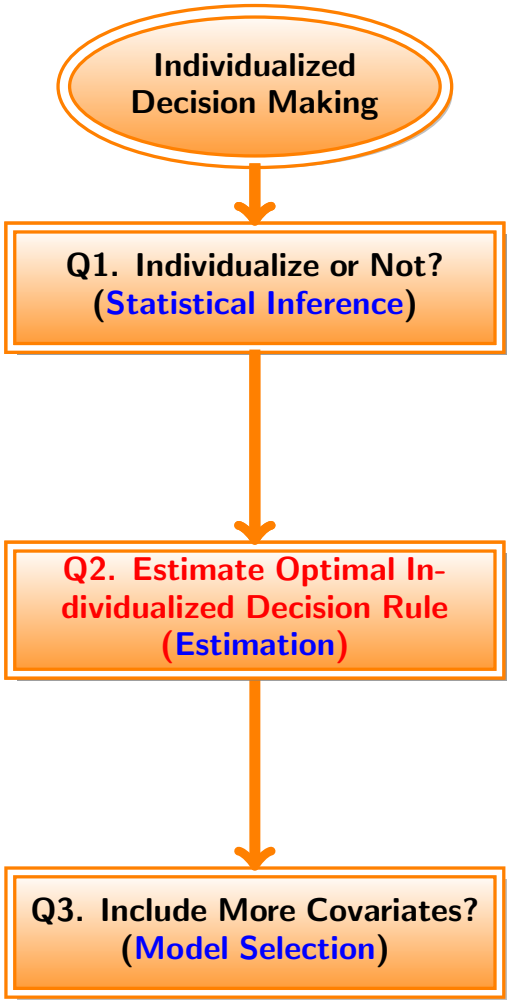
- Individualized decision making: Order dispatching strategy based on each order's characteristics:
  - starting point
  - destination
  - time
- Number of observations: 10^8 orders per week in Guangzhou (heat map at the right)

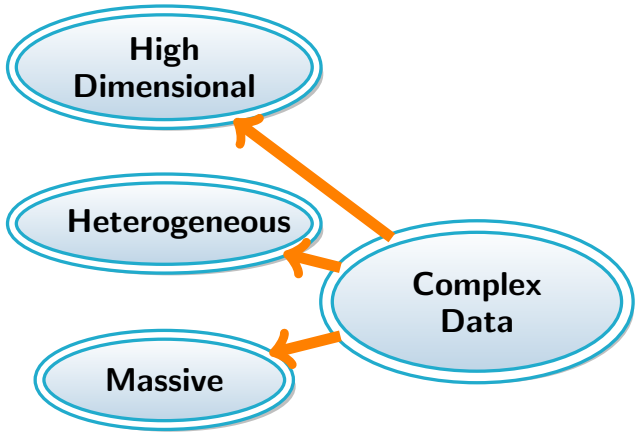


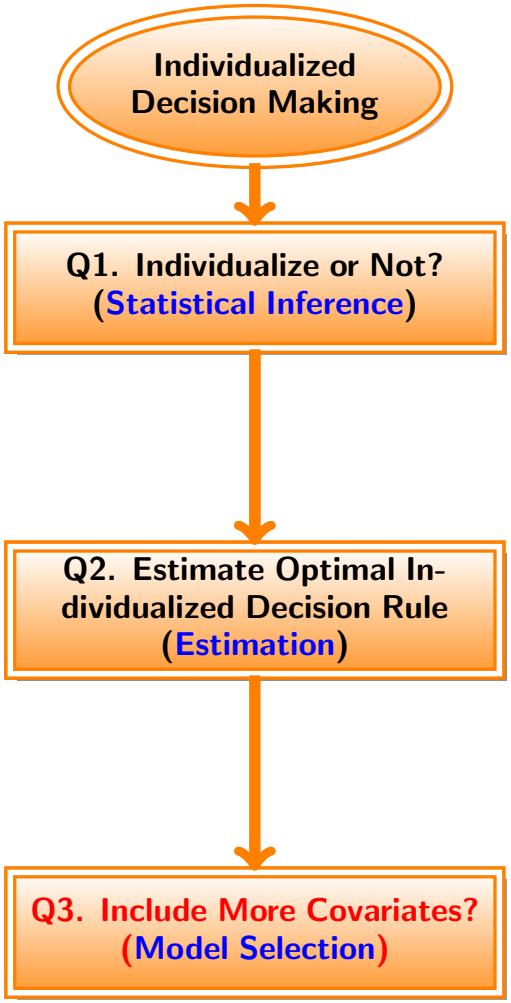


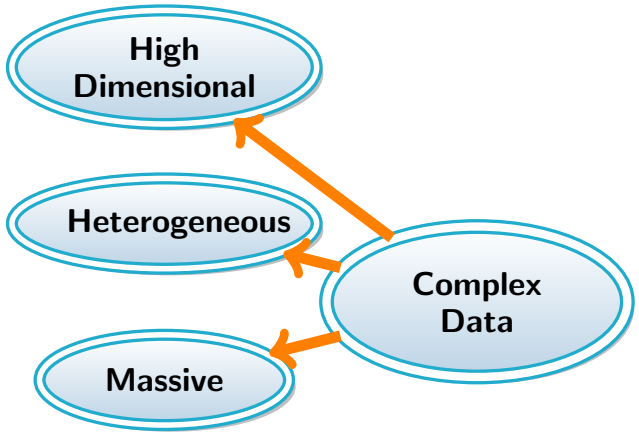

Individualized Decision Making
Q1. Individualize or Not? (Statistical Inference)
Q2. Estimate Optimal Individualized Decision Rule (Estimation)
Q3. Include More Covariates? (Model Selection)

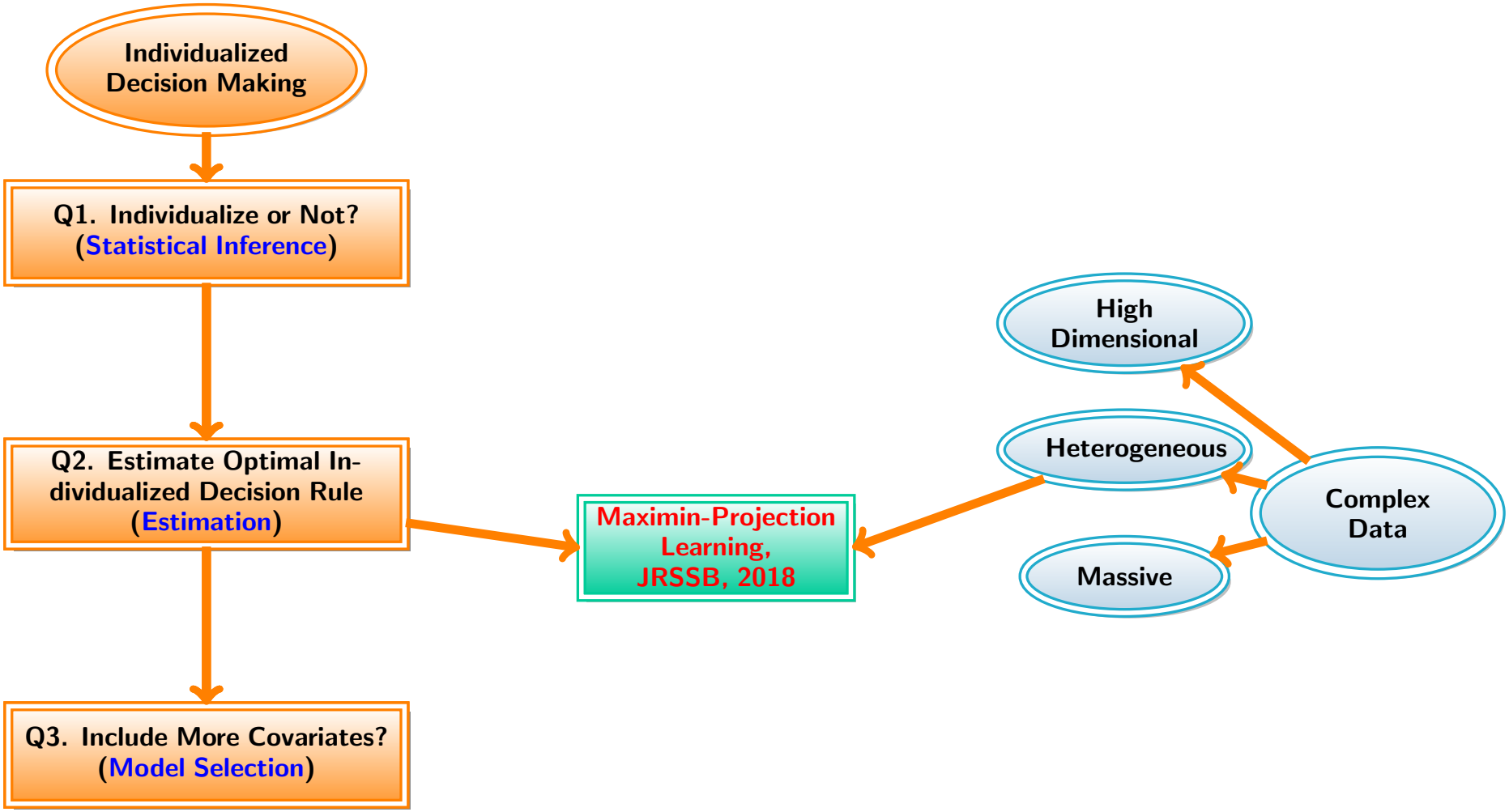

High Dimensional
Heterogeneous
Massive
Complex Data


Individualized Decision Making
Q1. Individualize or Not? (Statistical Inference)
Q2. Estimate Optimal Individualized Decision Rule (Estimation)
Q3. Include More Covariates? (Model Selection)


High Dimensional
Heterogeneous
Massive
Complex Data


Individualized Decision Making
Q1. Individualize or Not? (Statistical Inference)
Q2. Estimate Optimal Individualized Decision Rule (Estimation)
Q3. Include More Covariates? (Model Selection)


High Dimensional
Heterogeneous
Massive
Complex Data


Individualized Decision Making
Q1. Individualize or Not? (Statistical Inference)
Q2. Estimate Optimal Individualized Decision Rule (Estimation)
Q3. Include More Covariates? (Model Selection)
Maximin-Projection Learning, JRSSB, 2018
High Dimensional
Heterogeneous
Massive
Complex Data

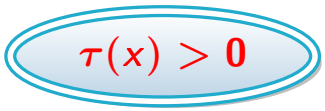
Maximin-Projection Learning for Individualized Decision Making with Heterogeneous Data

joint work with Rui Song, Wenbin Lu and Bo Fu

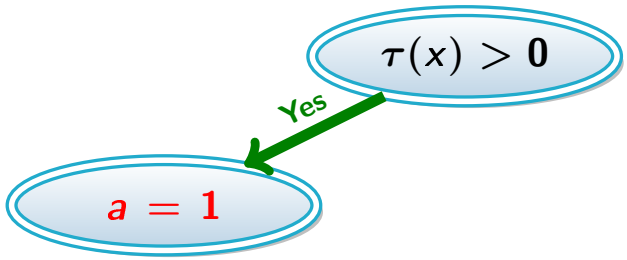
——*Journal of the Royal Statistical Society: Series B (2018)*, **80**, *681-702.*

- Individualized treatment regime (ITR): a function maps the covariate space to the space of available treatment options.
- Objective: identify the optimal ITR to reach the best 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: $\tau(x) = Q(1, x) - Q(0, x)$.

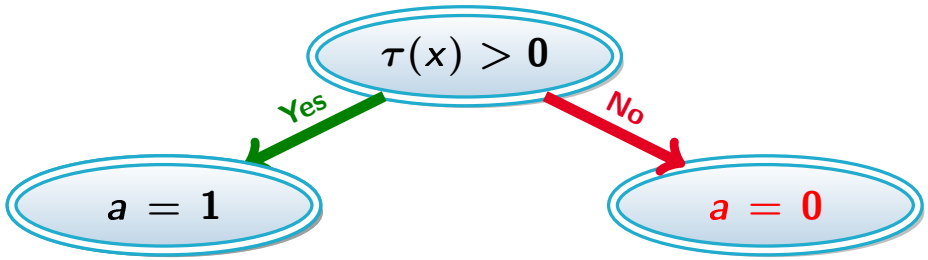
- Individualized treatment regime (ITR): a function maps the covariate space to the space of available treatment options.
- Objective: identify the optimal ITR to reach the best clinical outcome.
- Single stage study:
  - Covariates X: p-dimensional vector.
  - Treatment A: 1 for the treatment, 0 for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



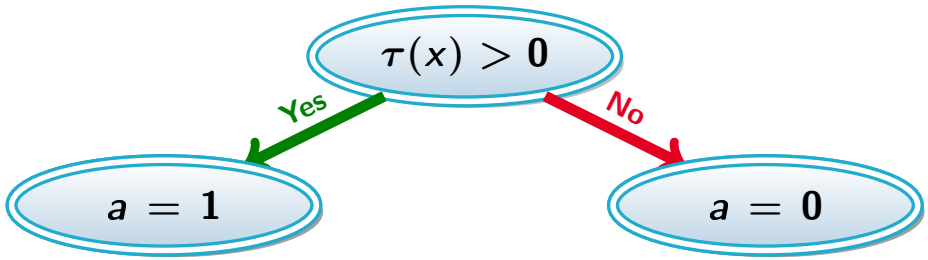
- Individualized treatment regime (ITR): a function maps the covariate space to the space of available treatment options.
- Objective: identify the optimal ITR to reach the best clinical outcome.
- Single stage study:
  - Covariates X: p-dimensional vector.
  - Treatment A: 1 for the treatment, 0 for 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.
- Objective: identify the optimal ITR to reach the best 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: $\tau(x) = Q(1, x) - Q(0, x)$.



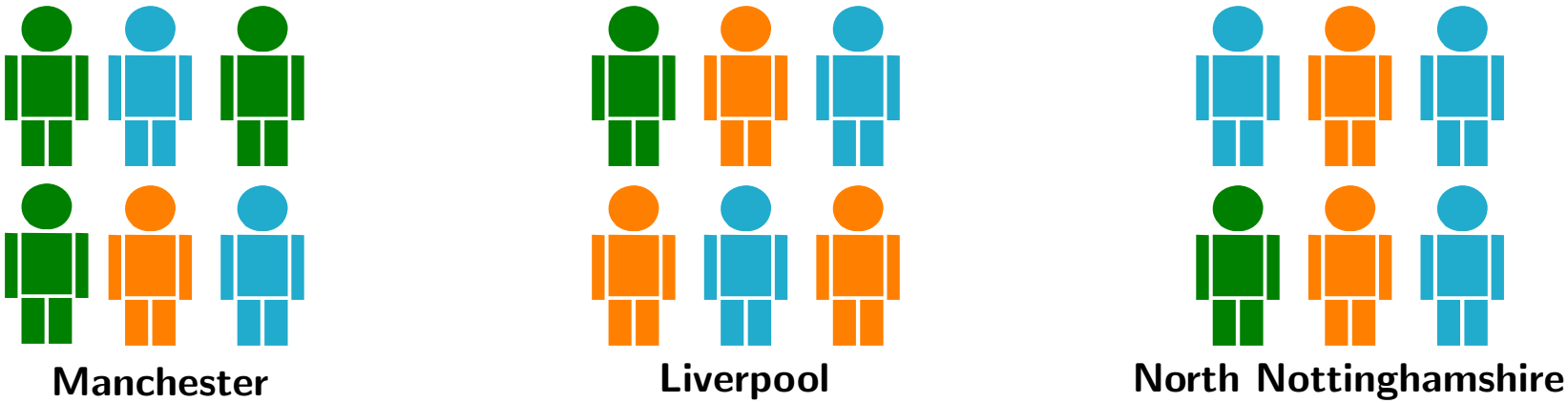
- Individualized treatment regime (ITR): a function maps the covariate space to the space of available treatment options.
- Objective: identify the optimal ITR to reach the best clinical outcome.
- Single stage study:
  - Covariates X: p-dimensional vector.
  - Treatment A: 1 for the treatment, 0 for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



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

Schizophrenia study (Tarrier et al., 2004)

- Over 400 patients with schizophrenia enrolled in 3 treatment centres in England



- Patients enrolled at different treatment centres show heterogeneity in response to treatments (Tarrier et al., 2004; Dunn and Bentall, 2007)
- Heterogeneity due to the differences in characteristics (unobserved or partially observed) of treatment setting across centres
- The optimal ITR can vary across different centres

Schizophrenia study (Tarrier et al., 2004)

Question

- How to derive a reliable ITR to the group of future patients (possibly from a new treatment centre)?
  - Any groupwise optimal ITR in the observed data might not be optimal (due to that characteristics explaining heterogeneity are unobserved)

Solution

- Provide an overall ITR by aggregating groupwise optimal ITRs estimated from the observed dataset
  - How to effectively combine groupwise ITRs?
  - What is a good criterion for combining ITRs?

Optimal ITR with a single population

- 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)

Optimal ITR with heterogeneous populations

- Why important? Multicentre studies are becoming more and more popular
- Our proposed ITR: accounts for heterogeneity due to different treatment settings across centres

Notation

- Patients are coming from G groups (treatment centres).
- For the g-th group (treatment centre),
  - Y_g: response (larger values are desirable).
  - A_g: treatment (binary).
  - X_g: covariates (standardized, $\mathrm{E}(X_g) = 0$, $\mathrm{cov}(X_g) = I$).

Model group heterogeneity

- Groupwise contrast function:

$$\begin{aligned} \tau_g(x) &= \mathrm{E}(Y_g|A_g = 1, X_g = x) - \mathrm{E}(Y_g|A_g = 0, X_g = x) \\ &= x^T\beta_g + c_g = \bar{x}^T\theta_g, \end{aligned}$$

  where $\bar{x} = (x^T, 1)^T$, $\theta_g = (\beta_g^T, c_g)^T$.

- Sources of heterogeneity:
  - c_g: Groupwise marginal treatment effect $\mathrm{E}\tau_g(X_g)$, the average treatment effect (ATE) of the g-th group
  - β_g: Groupwise individualized treatment effect.

Idea

- Groupwise Optimal ITR: $\mathbb{I}(\tau_g(x) > 0) = \mathbb{I}(\bar{x}^T \theta_g > 0)$
- Recommended ITR: $\mathbb{I}(\bar{x}^T \theta > 0)$ subject to $\|\theta\|_2 = 1$ that achieves some "optimality"

Idea

- Groupwise Optimal ITR: $\mathbb{I}(\tau_g(x) > 0) = \mathbb{I}(\bar{x}^T\theta_g > 0)$
- Recommended ITR: $\mathbb{I}(\bar{x}^T\theta > 0)$ subject to $\|\theta\|_2 = 1$ that achieves some "optimality"

How to define "optimality"

- Define the reward function $R_g(\theta)$ given the decision $\mathbb{I}(\bar{x}^T\theta > 0)$.
- Maximin effects (maximize the minimum reward)

$$\theta^M = \arg\max_{\|\theta\|_2=1} \min_{g\in\{1,\ldots,G\}} R_g(\theta).$$

- Why "maximin"?
  - Minimize the risk of the worst-case scenario
  - Minimax strategy in game theory (Wald, 1945)
  - Good performance in dealing with data heterogeneity (empirically & theoretically)

How to choose reward function

Maximin effects

$$\theta^M = \arg\max_{\|\theta\|_2=1} \min_{g\in\{1,\ldots,G\}} R_g(\theta).$$

Example (Value Difference (VD))

$$VD_g(\theta) = EY_g^*(d(X_g,\theta)) - EY_g^*(0),$$

where $d(X_g,\theta) = \mathbb{I}(\bar{X}_g^T\theta > 0)$.
The maximin effects

$$\theta_{(1)}^M = \arg\max_{\|\theta\|_2=1} \min_{g\in\{1,\ldots,G\}} VD_g(\theta).$$

How to choose reward function

Example (average Percentage of making Correct Decisions (PCD))

$$PCD_g(\theta) = 1 - E|\mathbb{I}(\bar{X}_g^T \theta_g > 0) - \mathbb{I}(\bar{X}_g^T \theta > 0)|,$$

where $\bar{X}_g = (X_g^T, 1)^T$.

The maximin effects

$$\theta_{(2)}^M = \arg\max_{\|\theta\|_2=1} \min_{g \in \{1,\ldots,G\}} PCD_g(\theta).$$

How to choose reward function

Pros

- PCD and value difference function are clinically important

Cons

- The empirical estimators for the value difference and PCD are non-smooth and non-concave functionals of θ.
- Computationally, the estimating procedure is difficult to implement.
- Theoretically, the convergence rates of these maximin estimators are slower than $O_p(n^{-1/2})$.

A toy example

$c_1 = \cdots = c_G = 0$, $X_1, \ldots, X_G \sim N(0, I_p)$. For any $\theta = (\beta^T, 0)^T$,

$$\mathrm{VD}_g(\theta) = \theta_g^T \theta / \sqrt{2\pi}.$$

A toy example

$c_1 = \cdots = c_G = 0$, $X_1, \ldots, X_G \sim N(0, I_p)$. For any $\theta = (\beta^T, 0)^T$,

$$\mathrm{VD}_g(\theta) = \theta_g^T \theta / \sqrt{2\pi}.$$

Maximin-projection learning

$$\theta^M = \arg\max_{\|\theta\|_2=1} \min_{g \in \{1,\ldots,G\}} \theta_g^T \theta.$$

- When $\|\theta_1\|_2 = \cdots = \|\theta_G\|_2$,

$$\theta^M = \arg\max_{\|\theta\|_2=1} \min_g \theta_g^T \theta / \|\theta_g\|_2 = \arg\min_{\|\theta\|_2=1} \max_g \angle(\theta_g, \theta).$$

Similar to the maximin correlation approach (Avi-Itzhak et al., 1995; Lee et al., 2016).

Statistical interpretation

Theorem (Equivalence of θ^M and $\theta^M_{(1)}$)

Assume X_g's (after groupwise standardization) are i.i.d. spherically distributed and $c_1 = c_2 = \cdots = c_G$. Then,

$$\theta^M = \arg\max_{\|\theta\|_2=1} \min_{g\in\{1,\ldots,G\}} VD_g(\theta).$$

Theorem (Equivalence of θ^M and $\theta^M_{(2)}$)

Assume X_g's (after groupwise standardization) are i.i.d. spherically distributed, $c_1 = c_2 = \cdots = c_G$, and all $\|\beta_g\|_2$'s are the same. Then,

$$\theta^M = \arg\max_{\|\theta\|_2=1} \min_{g\in\{1,\ldots,G\}} PCD_g(\theta).$$

Only need to focus on θ^M!

Geometric characterization

Example: $\beta_1 = (1,0)^T, \beta_2 = (\cos(60°), \sin(60°))^T, \beta_3 = (\cos(70°), \sin(70°))^T, \beta_4 = (0,1)^T$, $c_1 = c_2 = c_3 = c_4 = 0$.

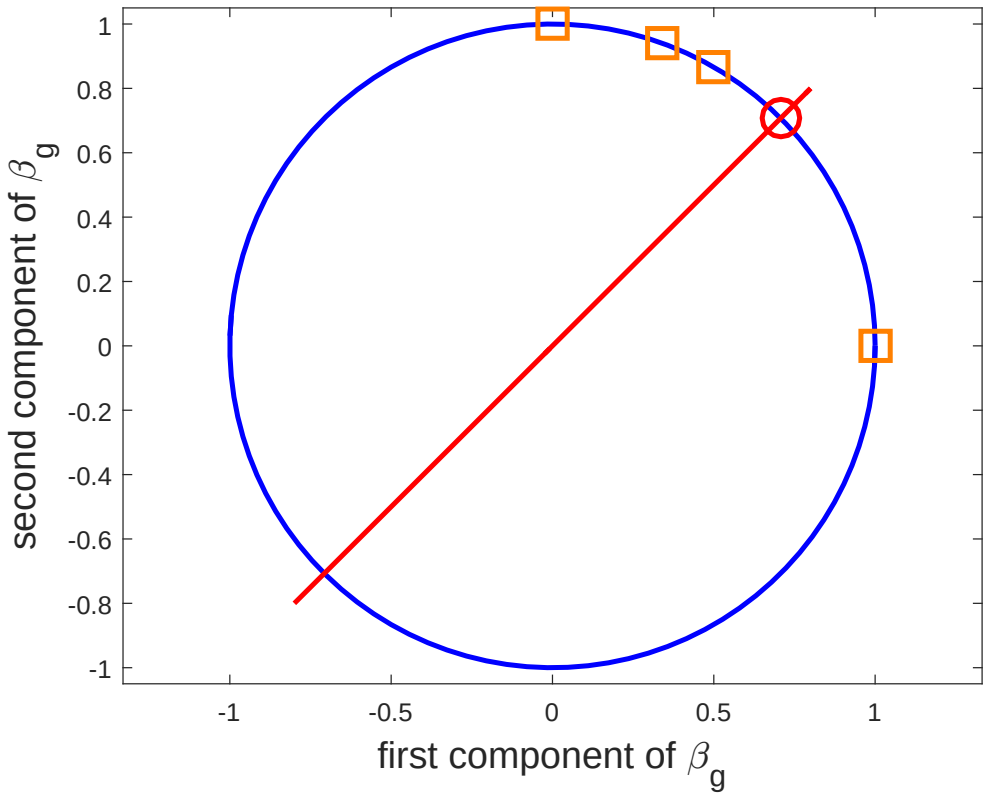


Figure: Plots of subgroup parameter β_g's (denoted by the square symbol) and the maximin effects (denoted by the circle symbol).

Geometric characterization

Example: $\beta_1 = (1,0)^T, \beta_2 = (\cos(60^\circ), \sin(60^\circ))^T, \beta_3 = (\cos(70^\circ), \sin(70^\circ))^T, \beta_4 = (0,1)^T$, $c_1 = c_2 = c_3 = c_4 = 0$.

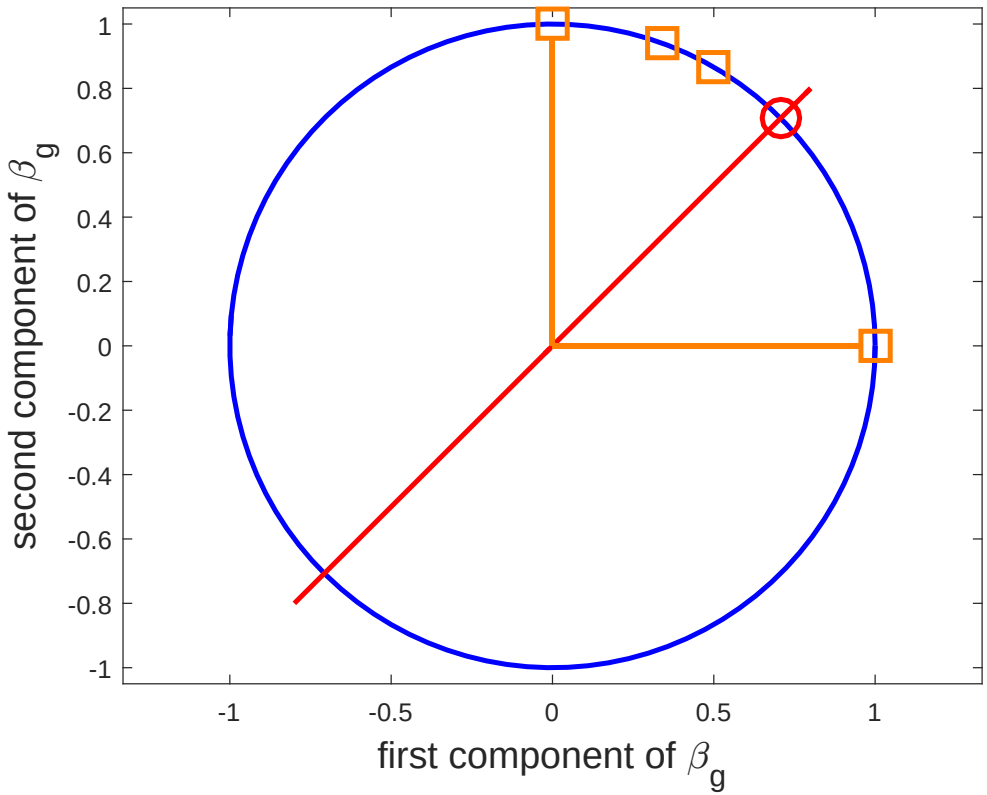


Figure: Plots of subgroup parameter β_g's (denoted by the square symbol) and the maximin effects (denoted by the circle symbol).

Equicorrelated point set (Avi-Itzhak et al., 1995)

Define the equicorrelated point set

$$\mathrm{E}(\Psi) = \left\{ t \in \mathbb{R}^m | t^T \Psi_j = t^T \Psi_i, \forall i, j \in \{1, \ldots, q\} \right\},$$

Ψ, an $m \times q$ matrix, and the optimal equicorrelated point

$$\mathrm{E}^*(\Psi) = \underset{\substack{t \in E(\Psi) \\ \|t\|_2 \leq 1}}{\arg\max} \left\{ t^T \Psi_i, \forall i \in \{1, \ldots, q\} \right\}.$$

Lemma

For any Ψ, *if* $\mathbf{1}_q \in C(\Psi^T)$, *then*

$$E^*(\Psi) = \{\mathbf{1}_q^T (\Psi^T \Psi)^+ \mathbf{1}_q\}^{-1/2} \Psi (\Psi^T \Psi)^+ \mathbf{1}_q.$$

Theorem

(I) Assume $F_0 = \max_{\|\theta\|_2=1} \min_g \theta^T \theta_g > 0$, *then there exists some set* $\mathcal{I}_0 \subseteq \{1, \ldots, G\}$ *such that*

$$\theta_g^T \theta^M = F_0, \forall g \in \mathcal{I}_0, \quad \theta_g^T \theta^M > F_0, \forall g \notin \mathcal{I}_0,$$

and $\theta^M = E^*(\Theta_{\mathcal{I}_0})$, *where* $\Theta = [\theta_1, \theta_2, \ldots, \theta_G]$.

Theorem

(I) Assume $F_0 = \max_{\|\theta\|_2=1} \min_g \theta^T \theta_g > 0$, *then there exists some set* $\mathcal{I}_0 \subseteq \{1, \ldots, G\}$ *such that*

$$\theta_g^T \theta^M = F_0, \forall g \in \mathcal{I}_0, \quad \theta_g^T \theta^M > F_0, \forall g \notin \mathcal{I}_0,$$

and $\theta^M = E^*(\Theta_{\mathcal{I}_0})$, *where* $\Theta = [\theta_1, \theta_2, \ldots, \theta_G]$.
(II) In addition, we have

$$\theta^M = \{\mathbf{1}_{|\mathcal{I}_0|}^T (\Theta_{\mathcal{I}_0}^T \Theta_{\mathcal{I}_0})^+ \mathbf{1}_{|\mathcal{I}_0|}\}^{-1/2} \Theta_{\mathcal{I}_0} (\Theta_{\mathcal{I}_0}^T \Theta_{\mathcal{I}_0})^+ \mathbf{1}_{|\mathcal{I}_0|}.$$

Theorem

(I) Assume $F_0 = \max_{\|\theta\|_2=1} \min_g \theta^T \theta_g > 0$, *then there exists some set* $\mathcal{I}_0 \subseteq \{1, \ldots, G\}$ *such that*

$$\theta_g^T \theta^M = F_0, \forall g \in \mathcal{I}_0, \quad \theta_g^T \theta^M > F_0, \forall g \notin \mathcal{I}_0,$$

and $\theta^M = E^*(\Theta_{\mathcal{I}_0})$, *where* $\Theta = [\theta_1, \theta_2, \ldots, \theta_G]$.
(II) In addition, we have

$$\theta^M = \{\mathbf{1}_{|\mathcal{I}_0|}^T (\Theta_{\mathcal{I}_0}^T \Theta_{\mathcal{I}_0})^+ \mathbf{1}_{|\mathcal{I}_0|}\}^{-1/2} \Theta_{\mathcal{I}_0} (\Theta_{\mathcal{I}_0}^T \Theta_{\mathcal{I}_0})^+ \mathbf{1}_{|\mathcal{I}_0|}.$$

(III) Moreover, if the set of vectors $\theta_g, g \in \mathcal{I}_0$ *is linearly independent, then a necessary and sufficient condition for* $\theta^M = E^*(\Theta_{\mathcal{I}_0})$ *is that each element in* $(\Theta_{\mathcal{I}_0}^T \Theta_{\mathcal{I}_0})^{-1} \mathbf{1}_{|\mathcal{I}_0|}$ *is nonnegative. (Essential to establish statistical properties of the maximin estimator)*

Estimating procedure

- Assume estimators $\hat{\beta}_1, \ldots, \hat{\beta}_G, \hat{c}_0$ (computed based on Q-learning or A-learning) are available. Let $\hat{\theta}_g = (\hat{\beta}_g^T, \hat{c}_0)^T$.

$$\hat{\theta}^M = \arg\max_{\|\theta\|_2=1} \min_{g\in\{1,\ldots,G\}} \hat{\theta}_g^T \theta = \arg\max_{\|(\beta^T,c)\|_2=1} \min_{g\in\{1,\ldots,G\}} (\hat{\beta}_g^T \beta + \hat{c}_0 c).$$

- Step 1: Concave optimization problem

$$\hat{\beta}_0 = \arg\max_{\|\beta\|_2\le 1} \min_{g\in\{1,\ldots,G\}} \hat{\beta}_g^T \beta.$$

  - Lagrange dual problem (QP, Lee et al., 2016):

$$\hat{\boldsymbol{x}} = \arg\min \boldsymbol{x}^T \widehat{B}^T \widehat{B} \boldsymbol{x}, \quad \text{subject to } \mathbf{1}_p^T \boldsymbol{x} = 1, \boldsymbol{x} \succeq 0,$$

    where $\widehat{B} = [\hat{\beta}_1, \ldots, \hat{\beta}_G]$.
  - $\hat{\beta}_0 = \widehat{B}\hat{\boldsymbol{x}}$. When $\|\hat{\beta}_0\|_2 > 0$, set $\hat{\beta}_0 = \hat{\beta}_0 / \|\hat{\beta}_0\|_2$.

Estimating procedure

- Assume estimators $\hat{\beta}_1, \ldots, \hat{\beta}_G, \hat{c}_0$ (computed based on Q-learning or A-learning) are available. Let $\hat{\theta}_g = (\hat{\beta}_g^T, \hat{c}_0)^T$.

$$\hat{\theta}^M = \arg\max_{\|\theta\|_2=1} \min_{g\in\{1,\ldots,G\}} \hat{\theta}_g^T \theta = \arg\max_{\|(\beta^T,c)\|_2=1} \min_{g\in\{1,\ldots,G\}} (\hat{\beta}_g^T \beta + \hat{c}_0 c).$$

- Step 2: Set $\kappa_0 = \min_{g\in\{1,\ldots,G\}} \hat{\beta}_0^T \hat{\beta}_g$.
  - If $\kappa_0 > 0$, set

$$\hat{\theta}^M = \left(\frac{\kappa_0 \hat{\beta}_0^T}{\sqrt{\kappa_0^2 + \hat{c}_0^2}}, \frac{\hat{c}_0}{\sqrt{\kappa_0^2 + \hat{c}_0^2}} \right)^T.$$

  - Else, set

$$\hat{\theta}^M = (\mathbf{0}_p^T, \mathrm{sgn}(\hat{c}_0))^T.$$

Estimating procedure

- Assume estimators $\hat{\beta}_1, \ldots, \hat{\beta}_G, \hat{c}_0$ (computed based on Q-learning or A-learning) are available. Let $\hat{\theta}_g = (\hat{\beta}_g^T, \hat{c}_0)^T$.

$$\hat{\theta}^M = \arg\max_{\|\theta\|_2=1} \min_{g\in\{1,\ldots,G\}} \hat{\theta}_g^T \theta = \arg\max_{\|(\beta^T,c)\|_2=1} \min_{g\in\{1,\ldots,G\}} (\hat{\beta}_g^T \beta + \hat{c}_0 c).$$

- Step 2: Set $\kappa_0 = \min_{g\in\{1,\ldots,G\}} \hat{\beta}_0^T \hat{\beta}_g$.
  - If $\kappa_0 > 0$, set

$$\hat{\theta}^M = \left(\frac{\kappa_0 \hat{\beta}_0^T}{\sqrt{\kappa_0^2 + \hat{c}_0^2}}, \frac{\hat{c}_0}{\sqrt{\kappa_0^2 + \hat{c}_0^2}} \right)^T.$$

  - Else, set

$$\hat{\theta}^M = (\mathbf{0}_p^T, \mathrm{sgn}(\hat{c}_0))^T.$$

- Output $\hat{d}(x) = \mathbb{I}(\bar{x}^T \hat{\theta}^M > 0)$.

Statistical properties

Theorem (Consistency and rate of convergence)

Let $\widehat{\Theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_G)$. *Under certain conditions, we have with probability going to* 1 *that* $\hat{\theta}^M$ *equals*

$$\{\mathbf{1}_{|\mathcal{I}_0|}^T (\widehat{\Theta}_{\mathcal{I}_0}^T \widehat{\Theta}_{\mathcal{I}_0})^{-1} \mathbf{1}_{|\mathcal{I}_0|}\}^{-1/2} \widehat{\Theta}_{\mathcal{I}_0} (\widehat{\Theta}_{\mathcal{I}_0}^T \widehat{\Theta}_{\mathcal{I}_0})^{-1} \mathbf{1}_{|\mathcal{I}_0|}.$$

In addition,

$$\|\hat{\theta}^M - \theta^M\|_2 = O\left(\max_{g \in \mathcal{I}_0} \|\hat{\theta}_g - \theta_g\|_2\right).$$

Theorem (Asymptotic normality)

Assume for all subgroup estimators are jointly asymptotically normal. Then, $\sqrt{n}(\hat{\theta}^M - \theta^M)$ *is asymptotically normal.*

Simulation setting

- Four groups of patients, each generated according as

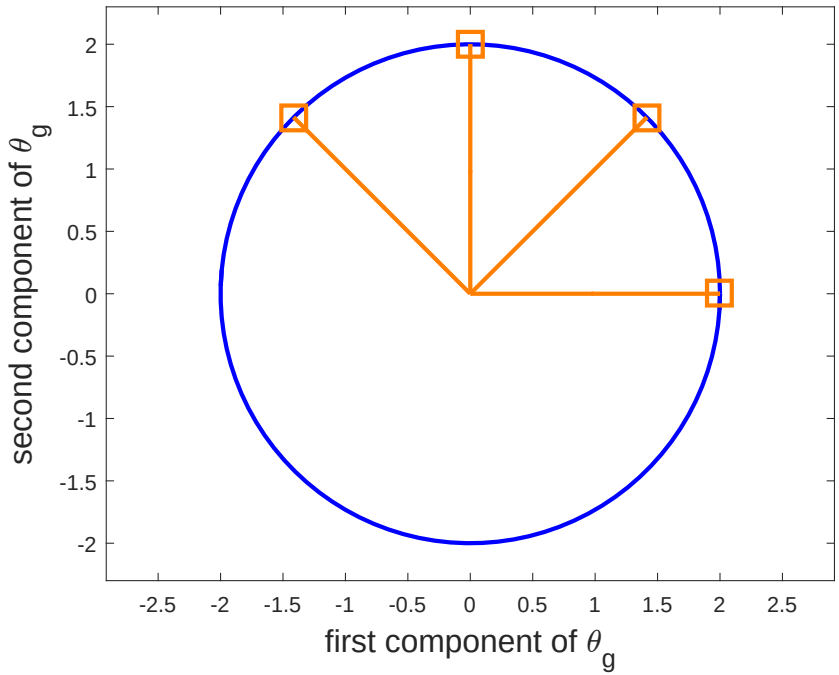
$$Y_{gj} = h(X_{gj}) + 2A_{gj}X_{gj}^T\beta_g + \varepsilon_{gj},$$

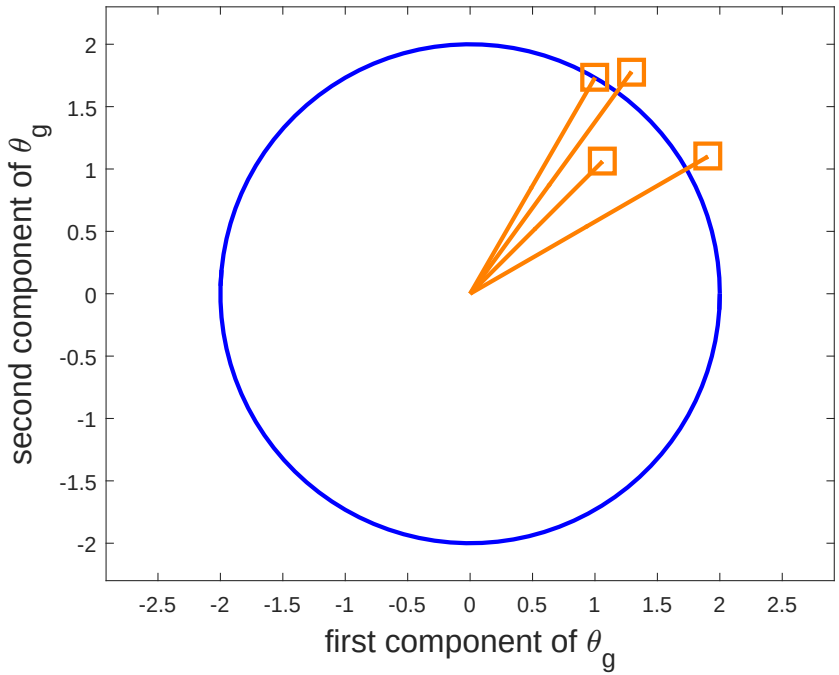
  $X_{gj} \overset{i.i.d}{\sim} N(0, I_4)$ and $\varepsilon_{gj} \overset{i.i.d}{\sim} N(0, 0.25)$.
- Two baseline models for h: linear and nonlinear.
- Two propensity score models for π: constant and probit.
- For each setting: subgroup estimator obtained using A-learning based on a linear model for h and logistic model for π:

1. S1: π correct, h correct,
2. S2: π correct, h wrong,
3. S3: π wrong, h correct,
4. S4: π wrong, h wrong.

Simulation setting (continued)

- Two scenarios for the subgroup parameters (representing different degrees of heterogeneity):
  - (I) same magnitudes, different angles
  - (II) similar angles, different magnitudes





Competing methods

- Random effects meta-analyses (DerSimonian and Laird, 1986; Jackson et al., 2010; Chen et al., 2012)
  - Compute $\hat{\theta}^R$ as a weighted average of $\hat{\theta}_g$'s,

$$\hat{\theta}^R = \left(\sum_{g=1}^{G} (\widehat{\Omega}_g + \widehat{\Omega}_0)^{-1} \right) \left(\sum_{g=1}^{G} (\widehat{\Omega}_g + \widehat{\Omega}_0)^{-1} \hat{\theta}_g \right).$$

  - Output $\mathbb{I}(\bar{x}^T \hat{\theta}^R > 0)$.
- Evaluation: leave-one-group-out cross-validation
  - Obtain ITR based any of the three training groups.
  - Evaluate its value difference on the remaining testing group.

Scenario I: same magnitudes, different angles

Table: **VD results** (with standard errors in parenthesis) under the estimated maximin ITR and the ITR obtained by random effects meta-analyses.

Testing group		First group	Second group	Third group	Fourth group
S1	**random**	-0.0002(0.001)	0.651(0.001)	0.651(0.001)	0.0005(0.001)
	maximin	0.002(0.001)	**0.736**(0.0004)	**0.736**(0.0004)	0.003(0.001)
S2	**random**	0.001(0.001)	0.650(0.001)	0.646(0.001)	-0.002(0.002)
	maximin	0.005(0.003)	**0.734**(0.001)	**0.734**(0.001)	0.002(0.002)
S3	**random**	-0.002(0.002)	0.648(0.002)	0.649(0.001)	0.001(0.002)
	maximin	-0.0003(0.002)	**0.734**(0.001)	**0.735**(0.001)	-0.003(0.002)
S4	**random**	-0.002(0.003)	0.638(0.003)	0.647(0.002)	0.008(0.003)
	maximin	-0.003(0.003)	**0.731**(0.001)	**0.728**(0.001)	-0.008(0.004)

Scenario I: same magnitudes, different angles

Table: **VD results** (with standard errors in parenthesis) under the estimated maximin ITR and the ITR obtained by random effects meta-analyses.

Testing group		First group	Second group	Third group	Fourth group
S1	**random**	-0.0002(0.001)	0.651(0.001)	0.651(0.001)	0.0005(0.001)
	maximin	0.002(0.001)	**0.736**(0.0004)	**0.736**(0.0004)	0.003(0.001)
S2	**random**	0.001(0.001)	0.650(0.001)	0.646(0.001)	-0.002(0.002)
	maximin	0.005(0.003)	**0.734**(0.001)	**0.734**(0.001)	0.002(0.002)
S3	**random**	-0.002(0.002)	0.648(0.002)	0.649(0.001)	0.001(0.002)
	maximin	-0.0003(0.002)	**0.734**(0.001)	**0.735**(0.001)	-0.003(0.002)
S4	**random**	-0.002(0.003)	0.638(0.003)	0.647(0.002)	0.008(0.003)
	maximin	-0.003(0.003)	**0.731**(0.001)	**0.728**(0.001)	-0.008(0.004)

Scenario I: same magnitudes, different angles

Table: **VD results** (with standard errors in parenthesis) under the estimated maximin ITR and the ITR obtained by random effects meta-analyses.

Testing group		First group	Second group	Third group	Fourth group
S1	**random**	-0.0002(0.001)	0.651(0.001)	0.651(0.001)	0.0005(0.001)
	maximin	0.002(0.001)	**0.736**(0.0004)	**0.736**(0.0004)	0.003(0.001)
S2	**random**	0.001(0.001)	0.650(0.001)	0.646(0.001)	-0.002(0.002)
	maximin	0.005(0.003)	**0.734**(0.001)	**0.734**(0.001)	0.002(0.002)
S3	**random**	-0.002(0.002)	0.648(0.002)	0.649(0.001)	0.001(0.002)
	maximin	-0.0003(0.002)	**0.734**(0.001)	**0.735**(0.001)	-0.003(0.002)
S4	**random**	-0.002(0.003)	0.638(0.003)	0.647(0.002)	0.008(0.003)
	maximin	-0.003(0.003)	**0.731**(0.001)	**0.728**(0.001)	-0.008(0.004)

Scenario II: similar angles, different magnitudes

Table: **VD results** (with standard errors in parenthesis) under the estimated maximin ITR and the ITR obtained by random effects meta-analyses.

Testing group		First group	Second group	Third group	Fourth group
S1	**random**	0.803(<0.001)	**0.598**(<0.001)	0.865(<0.001)	0.761(<0.001)
	maximin	**0.847**(<0.001)	0.588(<0.001)	0.865(<0.001)	0.769(<0.001)
S2	**random**	0.803(<0.001)	**0.598**(<0.001)	0.865(<0.001)	0.762(<0.001)
	maximin	**0.843**(0.001)	0.587(<0.001)	0.863(<0.001)	0.767(0.001)
S3	**random**	0.801(0.001)	**0.597**(<0.001)	0.864(<0.001)	0.761(0.001)
	maximin	**0.841**(0.001)	0.588(<0.001)	0.861(0.001)	0.765(0.001)
S4	**random**	0.804(0.001)	**0.597**(<0.001)	0.863(<0.001)	0.759(0.001)
	maximin	**0.826**(0.002)	0.587(0.001)	0.853(0.001)	0.756(0.002)

Scenario II: similar angles, different magnitudes

Table: **VD results** (with standard errors in parenthesis) under the estimated maximin ITR and the ITR obtained by random effects meta-analyses.

Testing group		First group	Second group	Third group	Fourth group
S1	**random**	0.803(<0.001)	**0.598**(<0.001)	0.865(<0.001)	0.761(<0.001)
	maximin	**0.847**(<0.001)	0.588(<0.001)	0.865(<0.001)	0.769(<0.001)
S2	**random**	0.803(<0.001)	**0.598**(<0.001)	0.865(<0.001)	0.762(<0.001)
	maximin	**0.843**(0.001)	0.587(<0.001)	0.863(<0.001)	0.767(0.001)
S3	**random**	0.801(0.001)	**0.597**(<0.001)	0.864(<0.001)	0.761(0.001)
	maximin	**0.841**(0.001)	0.588(<0.001)	0.861(0.001)	0.765(0.001)
S4	**random**	0.804(0.001)	**0.597**(<0.001)	0.863(<0.001)	0.759(0.001)
	maximin	**0.826**(0.002)	0.587(0.001)	0.853(0.001)	0.756(0.002)

Scenario II: similar angles, different magnitudes

Table: **VD results** (with standard errors in parenthesis) under the estimated maximin ITR and the ITR obtained by random effects meta-analyses.

Testing group		First group	Second group	Third group	Fourth group
S1	**random**	0.803(<0.001)	**0.598**(<0.001)	0.865(<0.001)	0.761(<0.001)
	maximin	**0.847**(<0.001)	0.588(<0.001)	0.865(<0.001)	0.769(<0.001)
S2	**random**	0.803(<0.001)	**0.598**(<0.001)	0.865(<0.001)	0.762(<0.001)
	maximin	**0.843**(0.001)	0.587(<0.001)	0.863(<0.001)	0.767(0.001)
S3	**random**	0.801(0.001)	**0.597**(<0.001)	0.864(<0.001)	0.761(0.001)
	maximin	**0.841**(0.001)	0.588(<0.001)	0.861(0.001)	0.765(0.001)
S4	**random**	0.804(0.001)	**0.597**(<0.001)	0.863(<0.001)	0.759(0.001)
	maximin	**0.826**(0.002)	0.587(0.001)	0.853(0.001)	0.756(0.002)

Scenario II: similar angles, different magnitudes

Table: **VD results** (with standard errors in parenthesis) under the estimated maximin ITR and the ITR obtained by random effects meta-analyses.

Testing group		First group	Second group	Third group	Fourth group
S1	**random**	0.803(<0.001)	**0.598**(<0.001)	0.865(<0.001)	0.761(<0.001)
	maximin	**0.847**(<0.001)	0.588(<0.001)	0.865(<0.001)	0.769(<0.001)
S2	**random**	0.803(<0.001)	**0.598**(<0.001)	0.865(<0.001)	0.762(<0.001)
	maximin	**0.843**(0.001)	0.587(<0.001)	0.863(<0.001)	0.767(0.001)
S3	**random**	0.801(0.001)	**0.597**(<0.001)	0.864(<0.001)	0.761(0.001)
	maximin	**0.841**(0.001)	0.588(<0.001)	0.861(0.001)	0.765(0.001)
S4	**random**	0.804(0.001)	**0.597**(<0.001)	0.863(<0.001)	0.759(0.001)
	maximin	**0.826**(0.002)	0.587(0.001)	0.853(0.001)	0.756(0.002)

Schizophrenia study

Table: ITRs based on maximin-projection learning and random effects meta-analyses, and their estimated value functions.

Testing group	Group 1		Group 2		Group 3	
	maximin	**random**	**maximin**	**random**	**maximin**	**random**
$\hat{\theta}_1$	0.10	-2.79	0.16	-2.89	0.25	0.11
$\hat{\theta}_2$	-0.90	-3.15	-0.01	-5.06	-0.003	4.62
$\hat{\theta}_3$	0.30	-1.25	0.70	3.26	0.68	1.05
$\hat{\mathrm{E}} Y_g^\star(d)$	26.25	25.33	29.91	32.04	**24.01**	**14.36**

Schizophrenia study

Table: ITRs based on maximin-projection learning and random effects meta-analyses, and their estimated value functions.

Testing group	Group 1		Group 2		Group 3	
	maximin	**random**	**maximin**	**random**	**maximin**	**random**
$\hat{\theta}_1$	0.10	-2.79	0.16	-2.89	0.25	0.11
$\hat{\theta}_2$	-0.90	-3.15	-0.01	-5.06	-0.003	4.62
$\hat{\theta}_3$	0.30	-1.25	0.70	3.26	0.68	1.05
$\hat{\mathrm{E}} Y_g^\star(d)$	26.25	25.33	29.91	32.04	**24.01**	**14.36**

Schizophrenia study

Table: ITRs based on maximin-projection learning and random effects meta-analyses, and their estimated value functions.

Testing group	Group 1		Group 2		Group 3	
	maximin	**random**	**maximin**	**random**	**maximin**	**random**
$\hat{\theta}_1$	0.10	-2.79	0.16	-2.89	0.25	0.11
$\hat{\theta}_2$	-0.90	-3.15	-0.01	-5.06	-0.003	4.62
$\hat{\theta}_3$	0.30	-1.25	0.70	3.26	0.68	1.05
$\hat{\mathrm{E}}Y_g^\star(d)$	26.25	25.33	29.91	32.04	**24.01**	**14.36**

Schizophrenia study

Table: ITRs based on maximin-projection learning and random effects meta-analyses, and their estimated value functions.

Testing group	Group 1		Group 2		Group 3	
	maximin	**random**	**maximin**	**random**	**maximin**	**random**
$\hat{\theta}_1$	0.10	-2.79	0.16	-2.89	0.25	0.11
$\hat{\theta}_2$	-0.90	-3.15	-0.01	-5.06	-0.003	4.62
$\hat{\theta}_3$	0.30	-1.25	0.70	3.26	0.68	1.05
$\hat{\mathrm{E}}Y_g^\star(d)$	26.25	25.33	29.91	32.04	**24.01**	**14.36**

Some thoughts on why "maximin" works better

Some thoughts on why "maximin" works better

- Random design setting: $\theta_1, \ldots, \theta_G \overset{iid}{\sim} F_\theta$
- Maximin effects

$$\theta^M = \underset{\|\theta\|_2=1}{\arg\max} \min_{g \in \{1,\ldots,G\}} \theta^T \theta_g$$

- Random effects meta-analyses

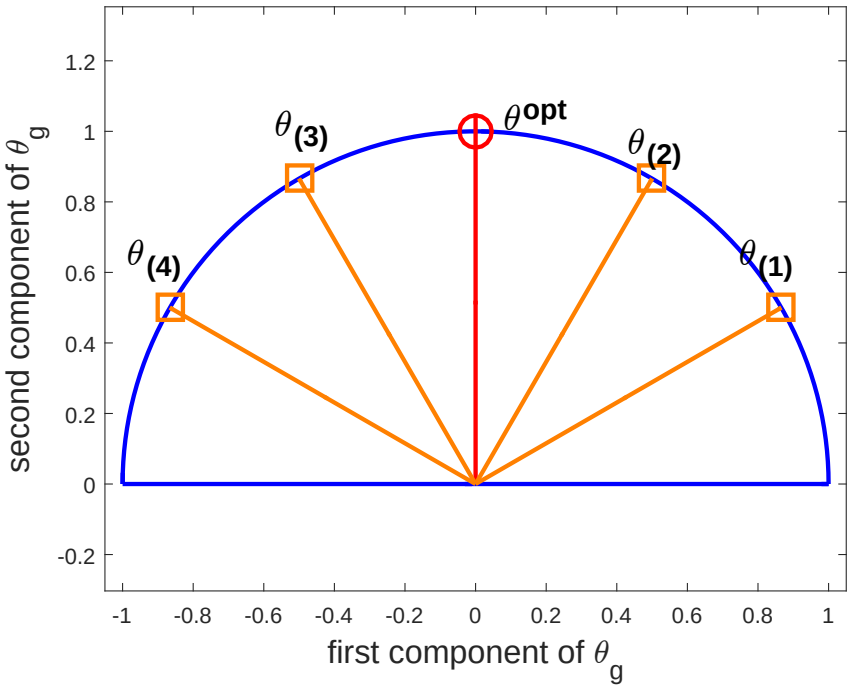
$$\hat{\theta}^R \xrightarrow{P} \frac{1}{G} \sum_{g=1}^{G} \theta_g \equiv \theta^R.$$

- Compare the expected value difference function with respect to the group of future patients

$$\mathrm{E}\{\mathrm{VD}(\theta_{G+1}; \theta^M) | \theta_1, \ldots, \theta_G, \theta_{G+1} \sim F_\theta\} \quad \text{v.s}$$

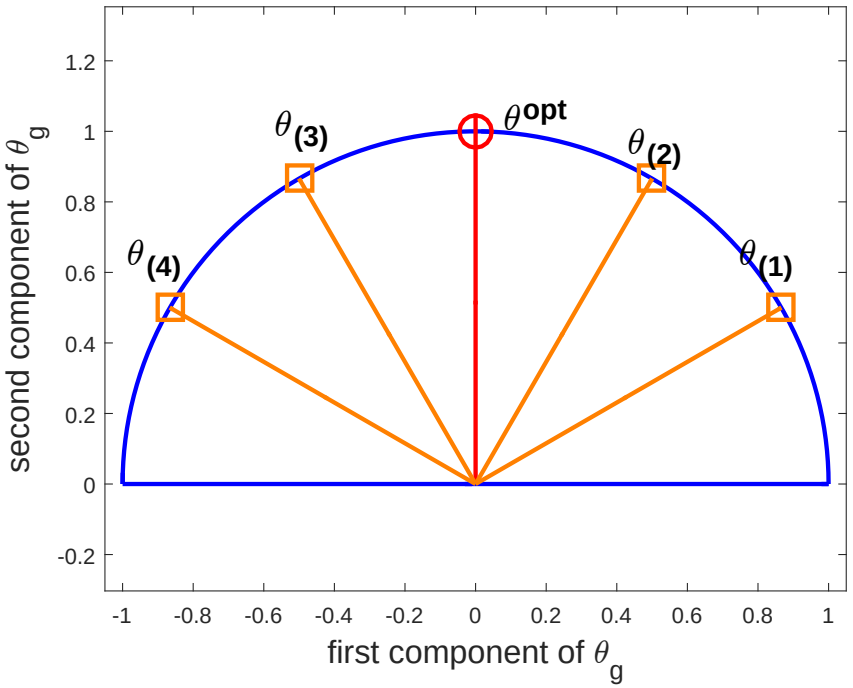
$$\mathrm{E}\{\mathrm{VD}(\theta_{G+1}; \theta^R) | \theta_1, \ldots, \theta_G, \theta_{G+1} \sim F_\theta\}$$

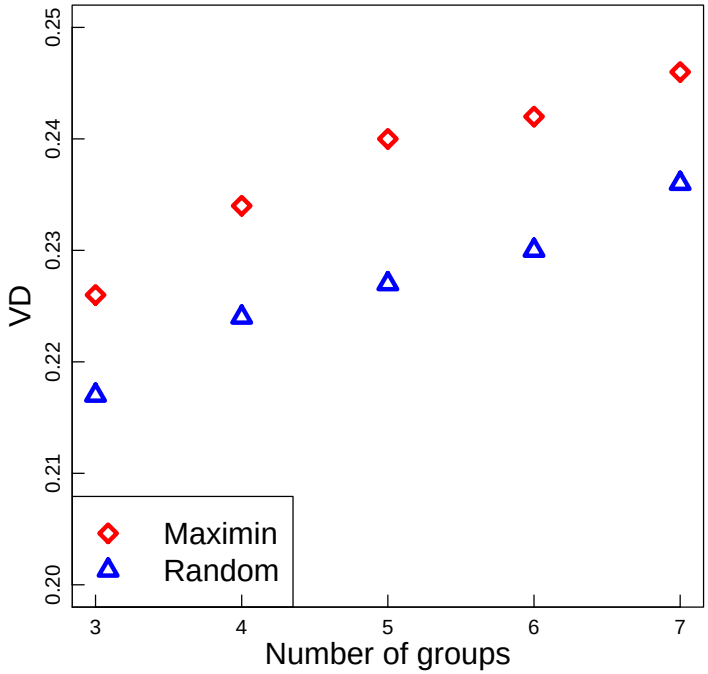
Toy example



- $\theta^{opt} = \arg\max_\theta \mathrm{E}\{\mathrm{VD}(\theta_{G+1};\theta)|\theta_{G+1} \sim F_\theta\} \propto \mathrm{E}(\theta|\theta \sim F_\theta)$
- $\theta^R = \sum_{g=1}^G \theta_g/G$, $\|\theta^R/\|\theta^R\|_2 - \theta^{opt}\|_2 = O_p(G^{-1/2})$
- $\theta^M \propto (\theta_{(G)} + \theta_{(1)})/2$, $\|\theta^M - \theta^{opt}\|_2 = O_p(G^{-1})$

Toy example





- $\theta^{opt} = \arg\max_\theta \mathrm{E}\{\mathrm{VD}(\theta_{G+1};\theta)|\theta_{G+1} \sim F_\theta\} \propto \mathrm{E}(\theta|\theta \sim F_\theta)$
- $\theta^R = \sum_{g=1}^G \theta_g/G$, $\|\theta^R/\|\theta^R\|_2 - \theta^{opt}\|_2 = O_p(G^{-1/2})$
- $\theta^M \propto (\theta_{(G)} + \theta_{(1)})/2$, $\|\theta^M - \theta^{opt}\|_2 = O_p(G^{-1})$

In general

Assume (A1)-(A3) hold. Then, we have $\theta^M - \theta^{opt} = O_p(G^{-1})$ where

$$\theta^{opt} = \frac{\mathrm{E}(\theta|\theta \sim F_\theta)}{\|\mathrm{E}(\theta|\theta \sim F_\theta)\|_2}.$$

(A1) $\min_{b \in \mathrm{supp}(F_\theta)} b^T \theta^{opt} > 0$.

(A2) $\forall b \in \mathrm{supp}(F_\theta)$, $2(b^T \theta^{opt})\theta^{opt} - b \in \mathrm{supp}(F_\theta)$.

(A3) There exist some finite vectors $\theta_{0,1}, \ldots, \theta_{0,K}$ that satisfy

- (i) $\min_{b \in \mathrm{supp}(F_\theta)} \theta^T b = \min_{k \in \{1,\ldots,K\}} \theta^T \theta_{0,k}$, $\forall \theta$ with $\|\theta\|_2 = 1$.
- (ii) For $1 \le k \le K$ and any sufficiently small $\varepsilon > 0$, we have $\Pr_{\theta \sim F_\theta}(d(\theta, \theta_{0,k}) \le \varepsilon) = O(\varepsilon)$.
- (iii) Let $\Theta = [\theta_{0,1}, \ldots, \theta_{0,K}]$ and $\mathcal{I}_0$ be some subset of $\{1, \ldots, K\}$ such that $\theta^{opt} = \mathrm{E}^*(\Theta_{\mathcal{I}_0})$. Assume vectors in $\Theta_{\mathcal{I}_0}$ are linearly independent, each element in $(\Theta_{\mathcal{I}_0}^T \Theta_{\mathcal{I}_0})^{-1} \mathbf{1}_{\mathcal{I}_0}$ is nonzero, and $\theta \in C(\Theta_{\mathcal{I}_0}), \forall \theta \in F_\theta$.

In general

Assume (A1)-(A3) hold. Then, we have $\theta^M - \theta^{opt} = O_p(G^{-1})$ where

$$\theta^{opt} = \frac{\mathrm{E}(\theta|\theta \sim F_\theta)}{\|\mathrm{E}(\theta|\theta \sim F_\theta)\|_2}.$$

(A1) $\min_{b \in \mathrm{supp}(F_\theta)} b^T \theta^{opt} > 0$.

(A2) $\forall b \in \mathrm{supp}(F_\theta)$, $2(b^T \theta^{opt})\theta^{opt} - b \in \mathrm{supp}(F_\theta)$.

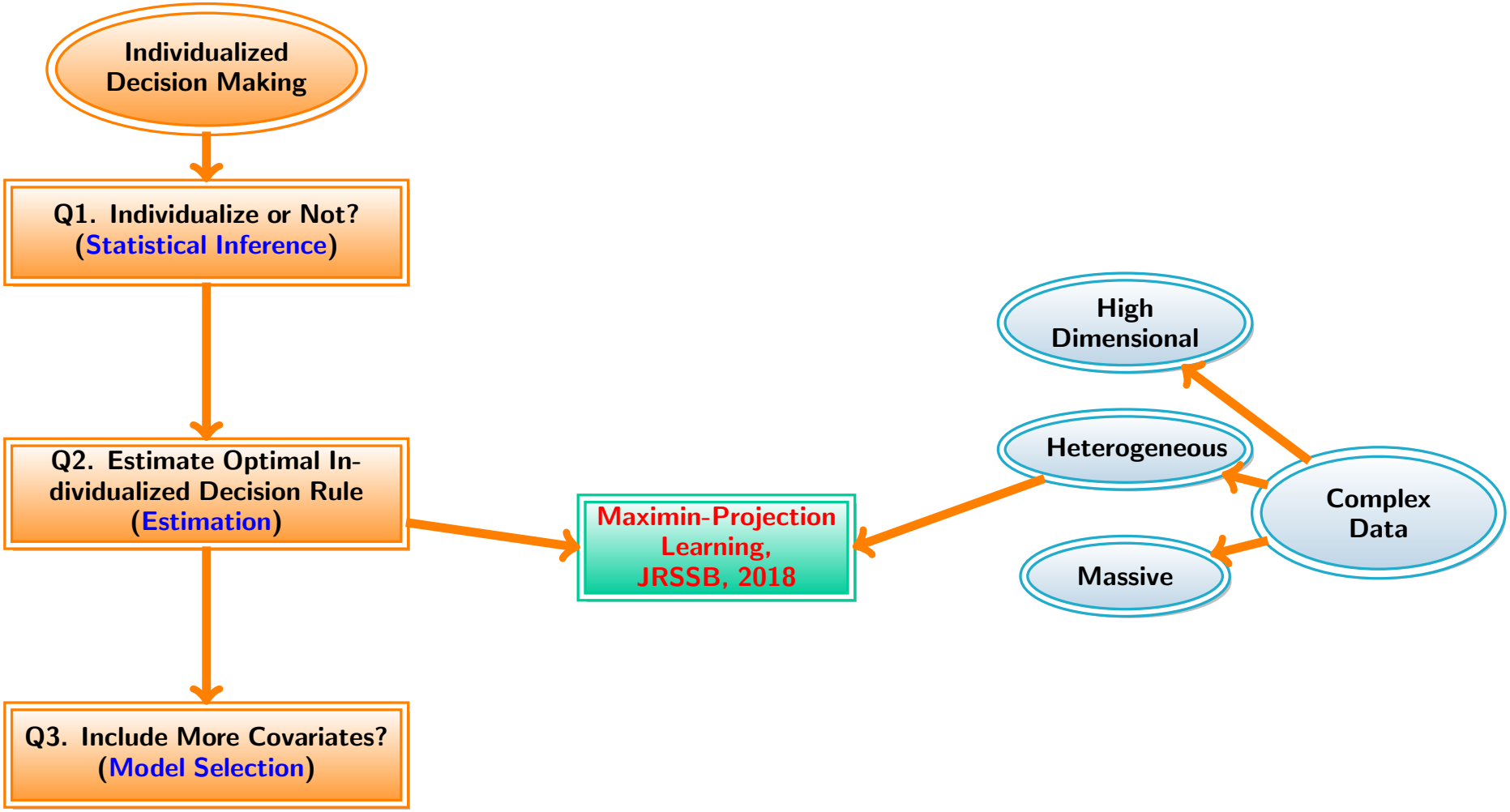
(A3) There exist some finite vectors $\theta_{0,1}, \ldots, \theta_{0,K}$ that satisfy

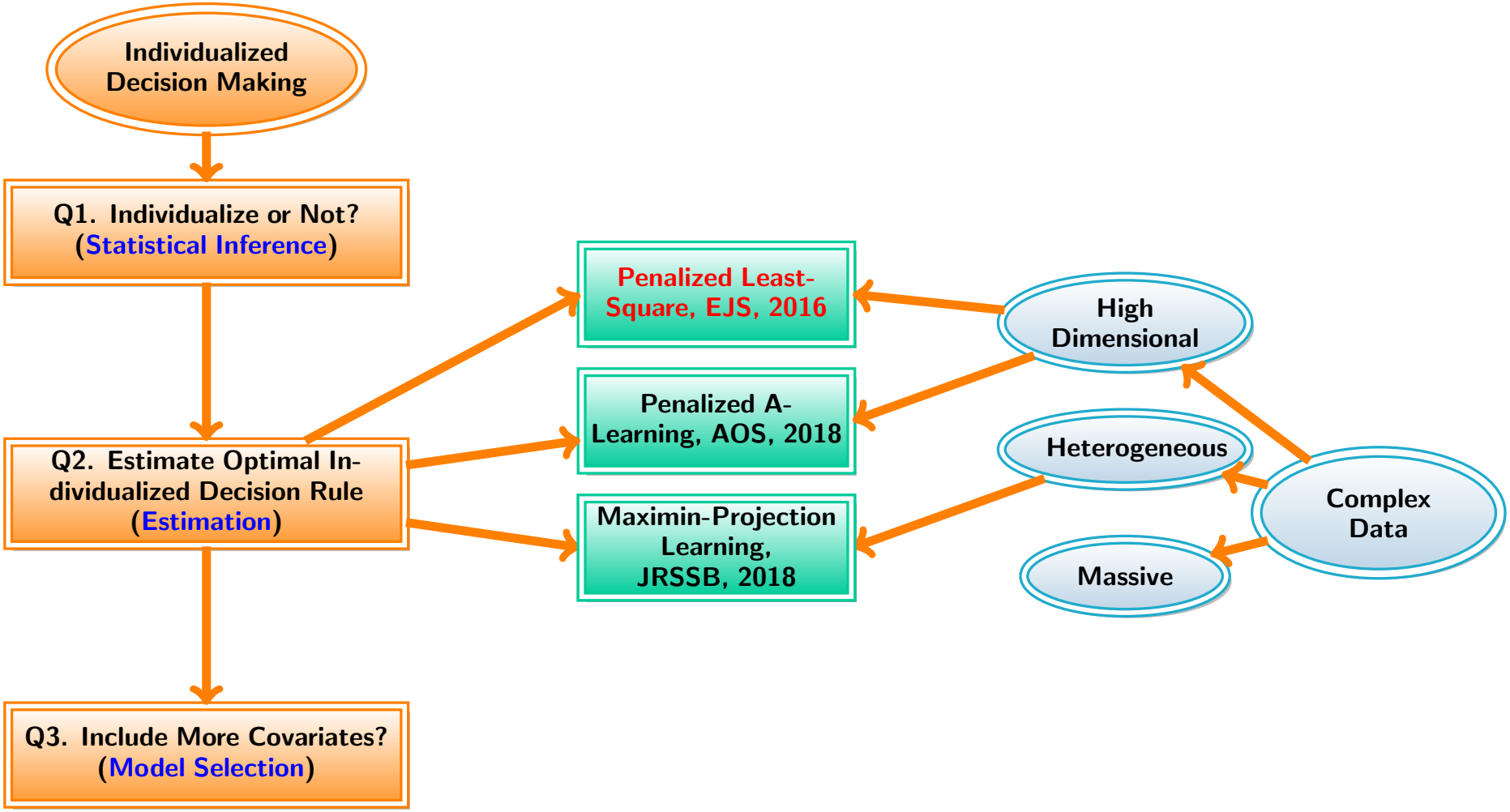
- (i) $\min_{b \in \mathrm{supp}(F_\theta)} \theta^T b = \min_{k \in \{1,\ldots,K\}} \theta^T \theta_{0,k}$, $\forall \theta$ with $\|\theta\|_2 = 1$.
- (ii) For $1 \le k \le K$ and any sufficiently small $\varepsilon > 0$, we have $\Pr_{\theta \sim F_\theta}(d(\theta, \theta_{0,k}) \le \varepsilon) = O(\varepsilon)$.
- (iii) Let $\Theta = [\theta_{0,1}, \ldots, \theta_{0,K}]$ and $\mathcal{I}_0$ be some subset of $\{1, \ldots, K\}$ such that $\theta^{opt} = \mathrm{E}^*(\Theta_{\mathcal{I}_0})$. Assume vectors in $\Theta_{\mathcal{I}_0}$ are linearly independent, each element in $(\Theta_{\mathcal{I}_0}^T \Theta_{\mathcal{I}_0})^{-1} \mathbf{1}_{\mathcal{I}_0}$ is nonzero, and $\theta \in C(\Theta_{\mathcal{I}_0}), \forall \theta \in F_\theta$.

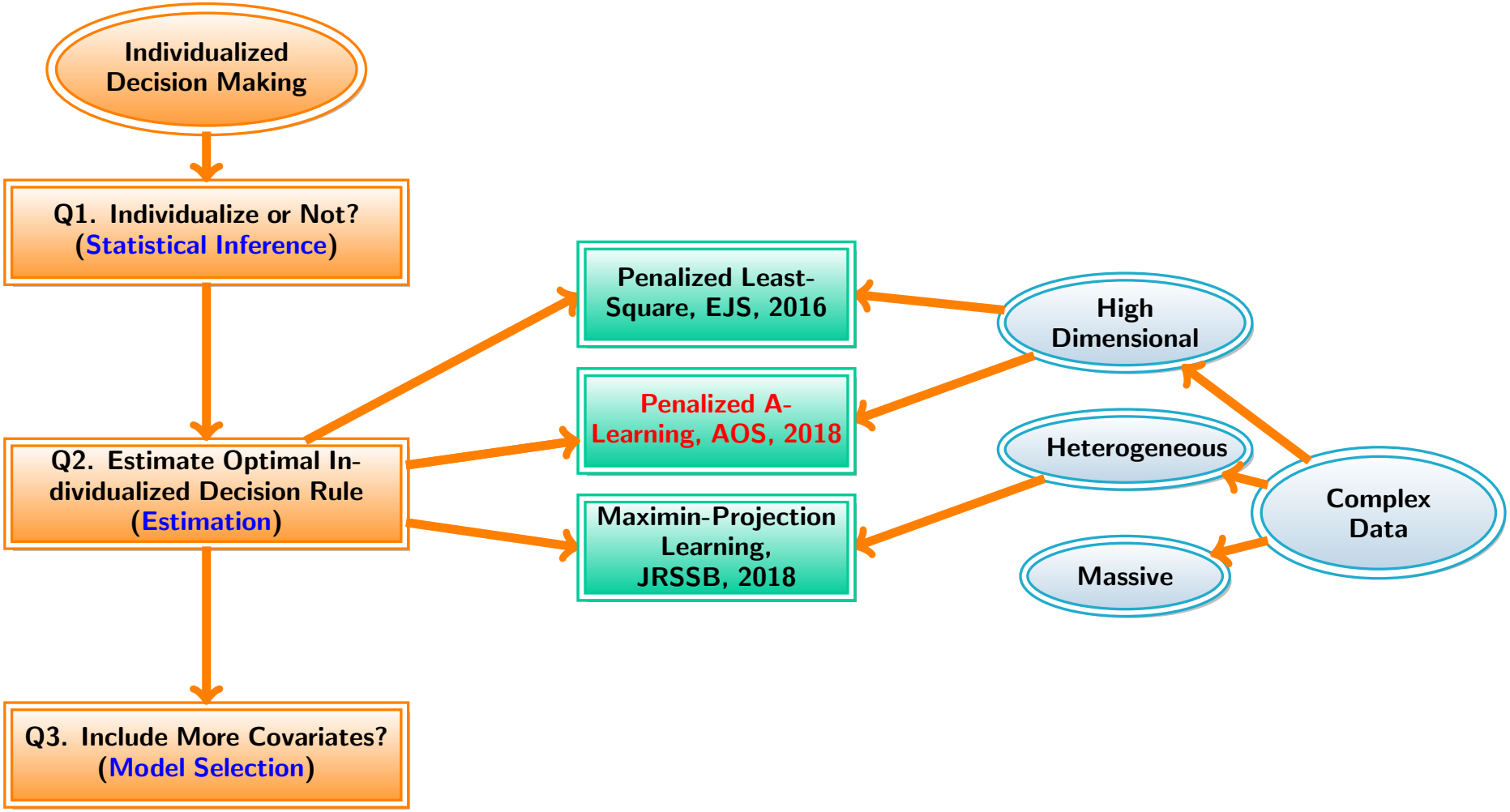
When (A2) is violated: a bias-variance trade-off

Summary

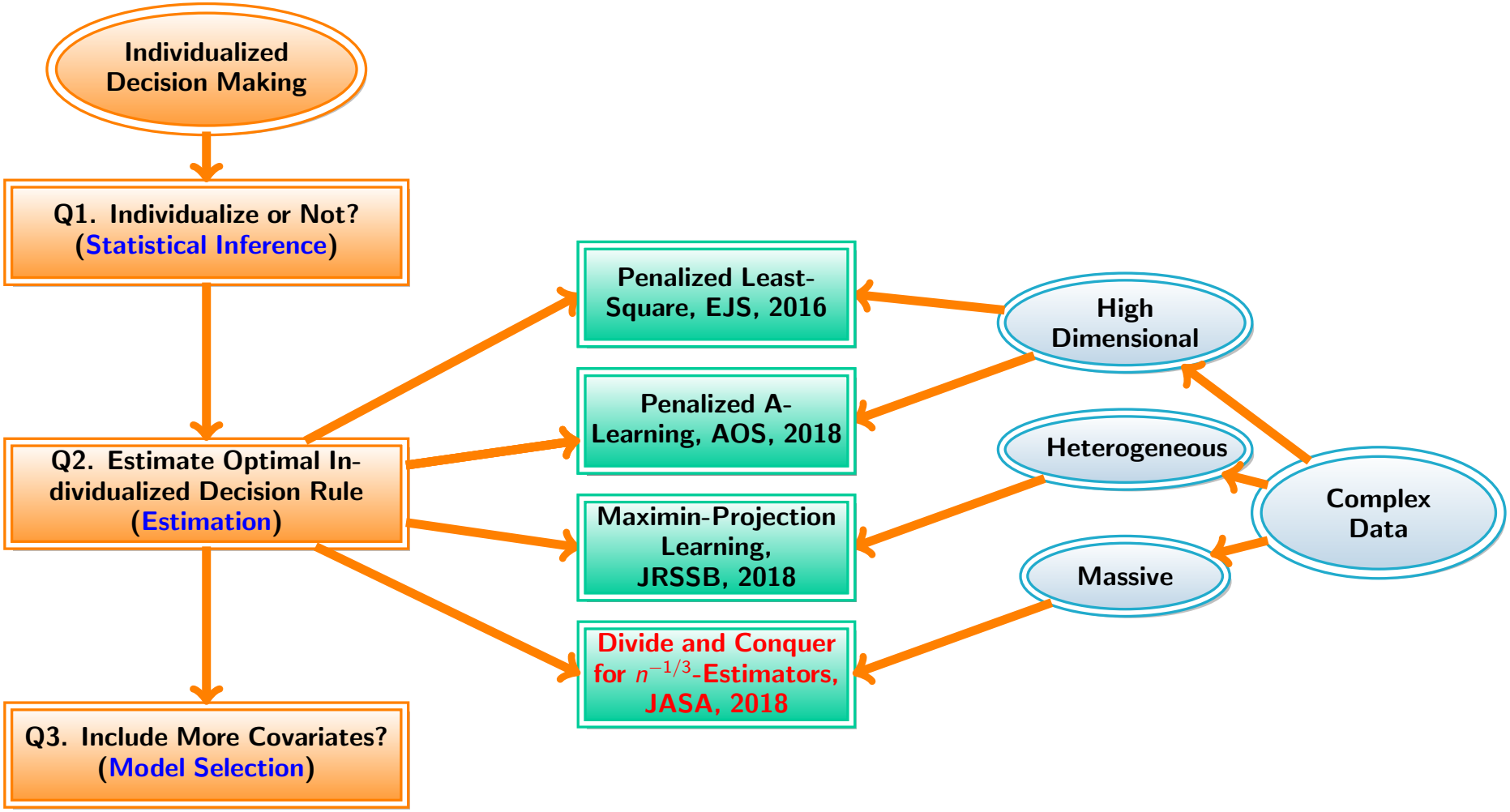
- Proposed a maximin-projection learning to derive a reliable ITR based on the observed data from different populations with heterogeneity in optimal individualized decision making
- Proposed maximin effects:
  - Meaningful interpretation (maximin value difference & PCD).
  - Nice geometric characterization (optimal equicorrelated point).
  - Efficient computation procedure (quadratic programming).
  - Appealing statistical properties (consistency & asymptotic normality).
  - Better performance (compared to random effects meta-analyses).

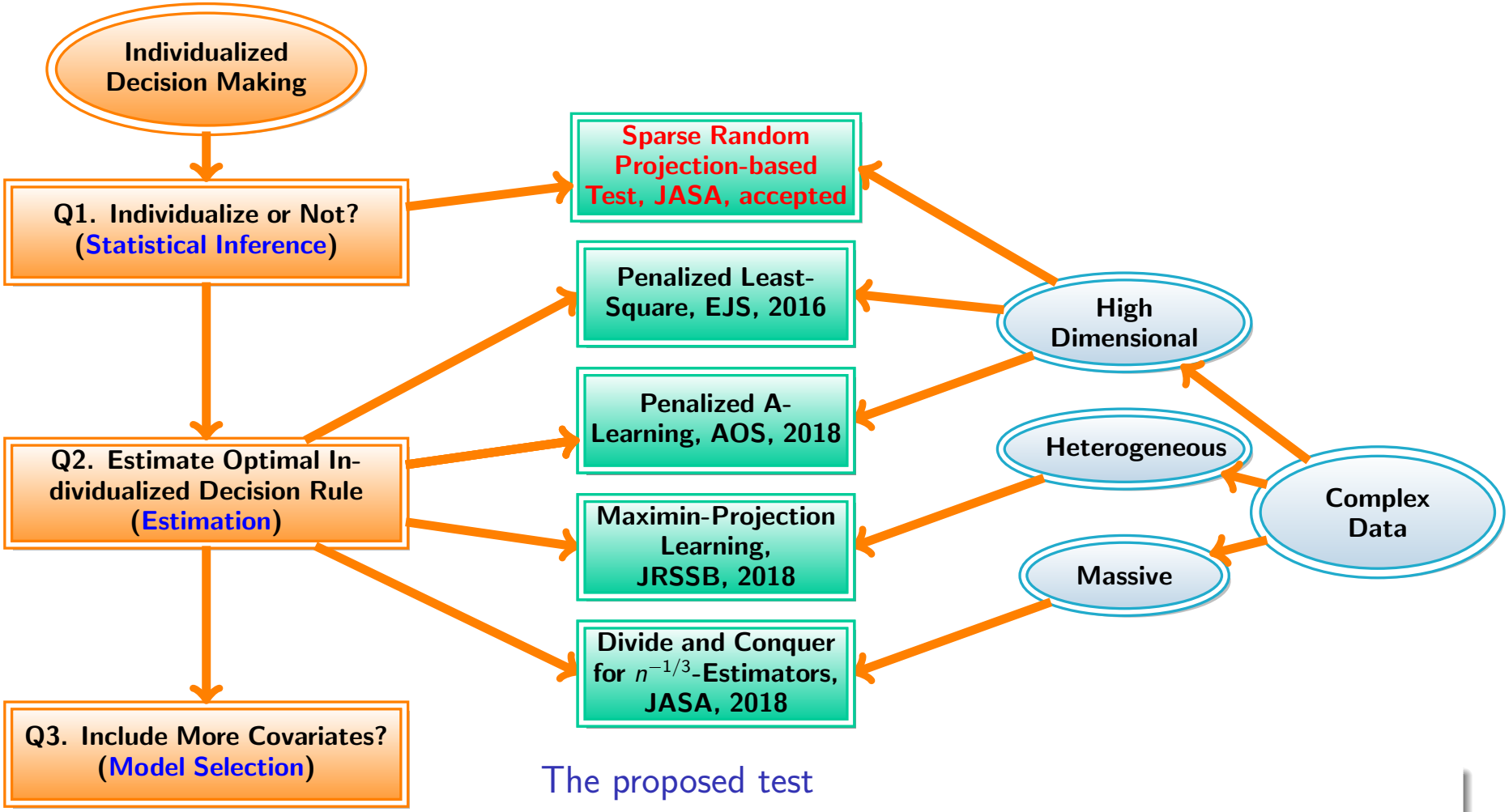

Individualized Decision Making
Q1. Individualize or Not? (Statistical Inference)
Q2. Estimate Optimal Individualized Decision Rule (Estimation)
Q3. Include More Covariates? (Model Selection)
Maximin-Projection Learning, JRSSB, 2018
High Dimensional
Heterogeneous
Massive
Complex Data


Individualized Decision Making
Q1. Individualize or Not? (Statistical Inference)
Q2. Estimate Optimal Individualized Decision Rule (Estimation)
Q3. Include More Covariates? (Model Selection)
Penalized Least-Square, EJS, 2016
Penalized A-Learning, AOS, 2018
Maximin-Projection Learning, JRSSB, 2018
High Dimensional
Heterogeneous
Massive
Complex Data



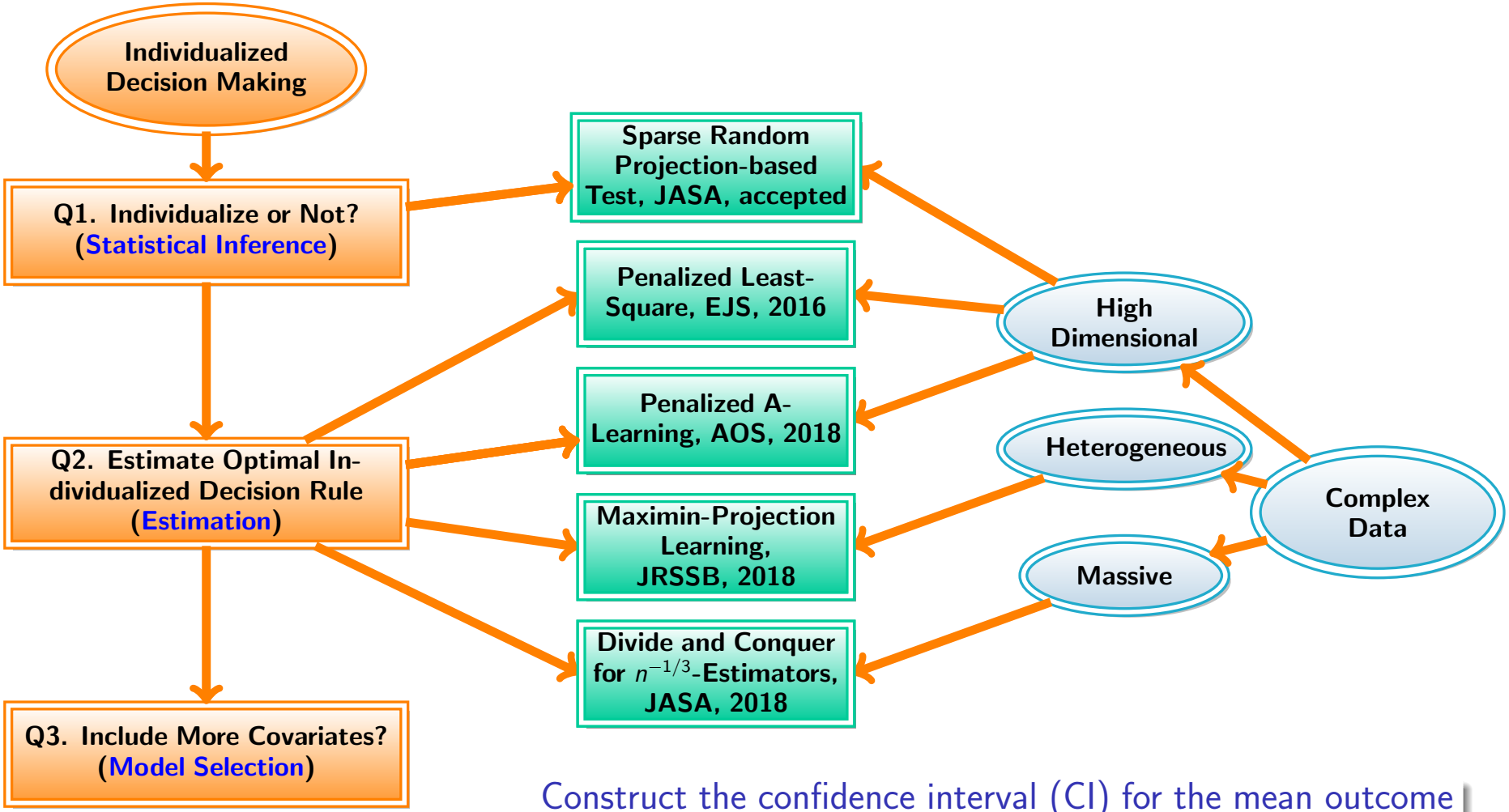
- PAL preserves the doubly-robustness property of classical A-learning when $p \gg n$


Individualized Decision Making
Q1. Individualize or Not? (Statistical Inference)
Q2. Estimate Optimal Individualized Decision Rule (Estimation)
Q3. Include More Covariates? (Model Selection)
Penalized Least-Square, EJS, 2016
Penalized A-Learning, AOS, 2018
Maximin-Projection Learning, JRSSB, 2018
Divide and Conquer for $n^{-1/3}$-Estimators, JASA, 2018
High Dimensional
Heterogeneous
Massive
Complex Data



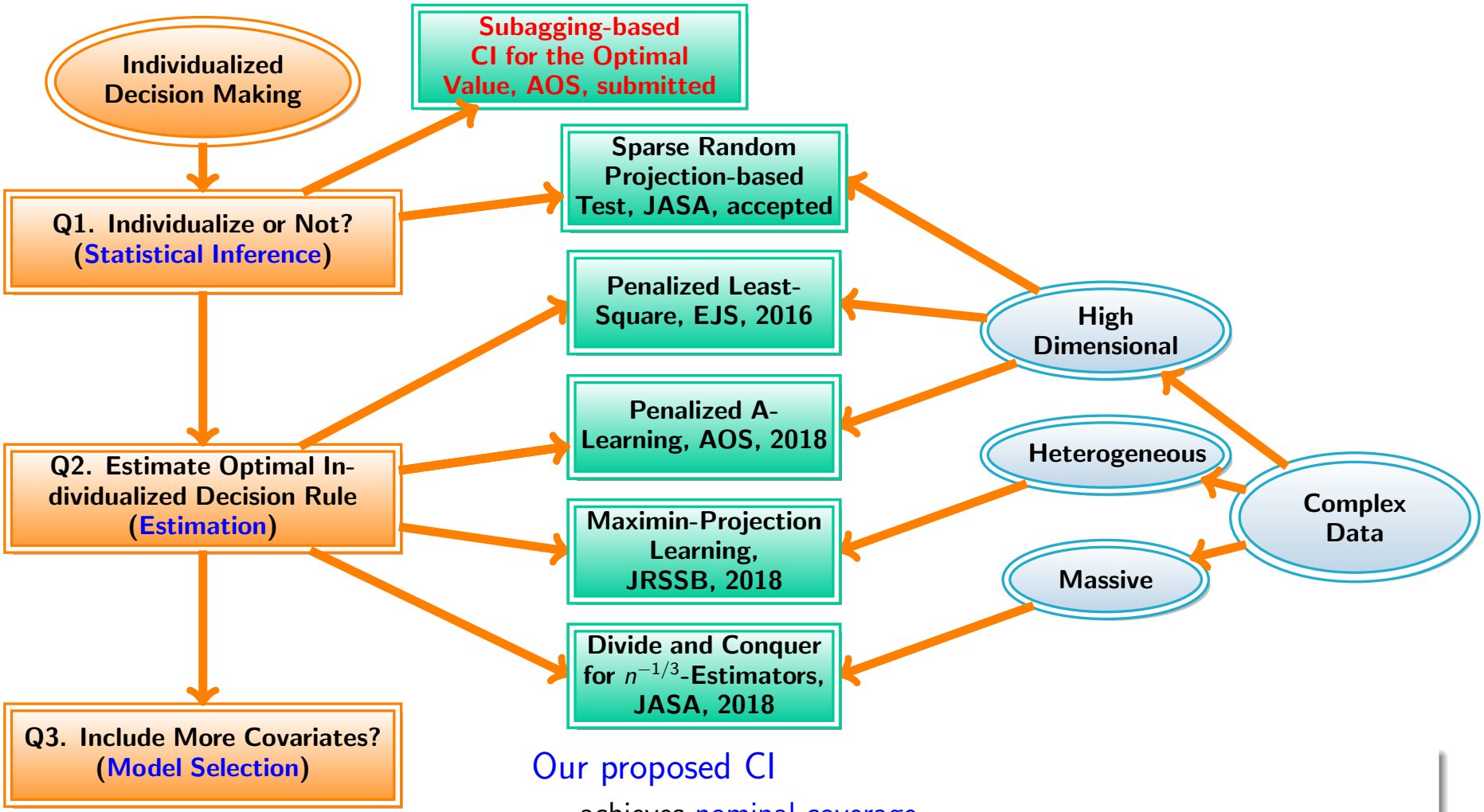
The proposed test

- H_0 : implementing the optimal individualized decision rule is equivalent to the "one-size-fits-all" method
- is constructed based on sparse random projections of covariates
- has the same asymptotic power function as the "oracle" test based on the "optimal" projection matrix



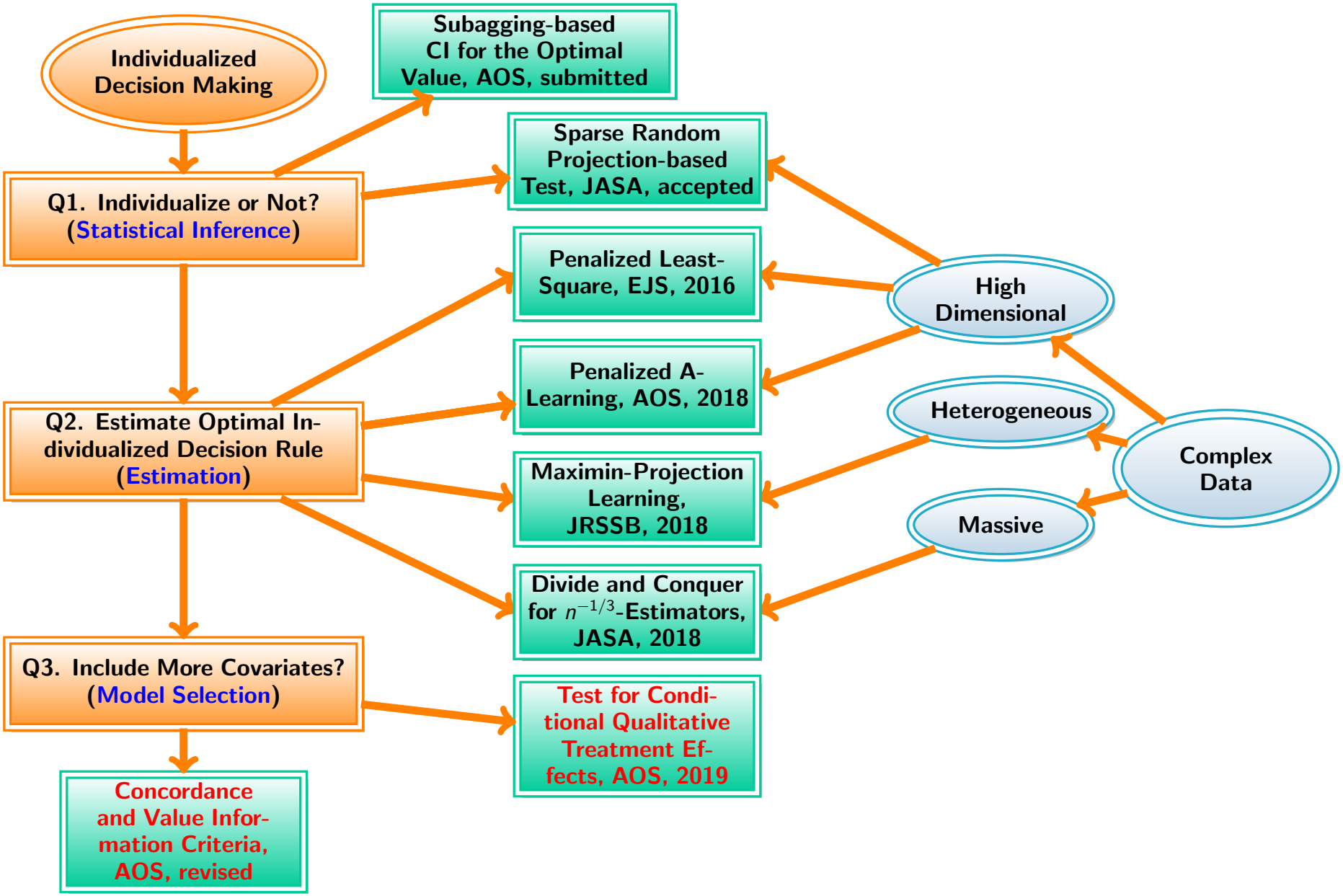
Construct the confidence interval (CI) for the mean outcome under an optimal individualized decision rule (optimal value function)

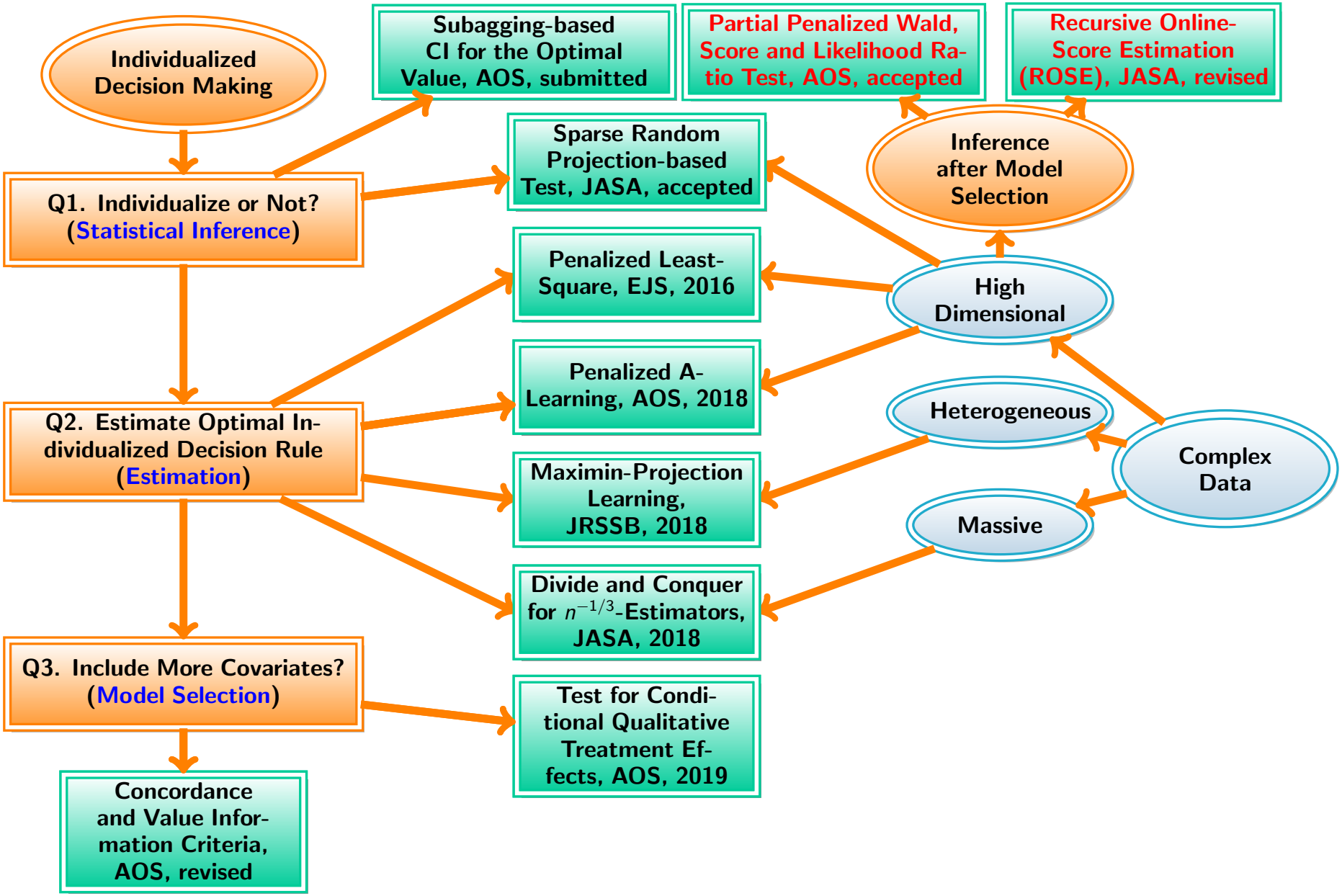
- has shown to be difficult in the nonregular cases (Robins, 2004; Robins and Rotnitzky, 2014).

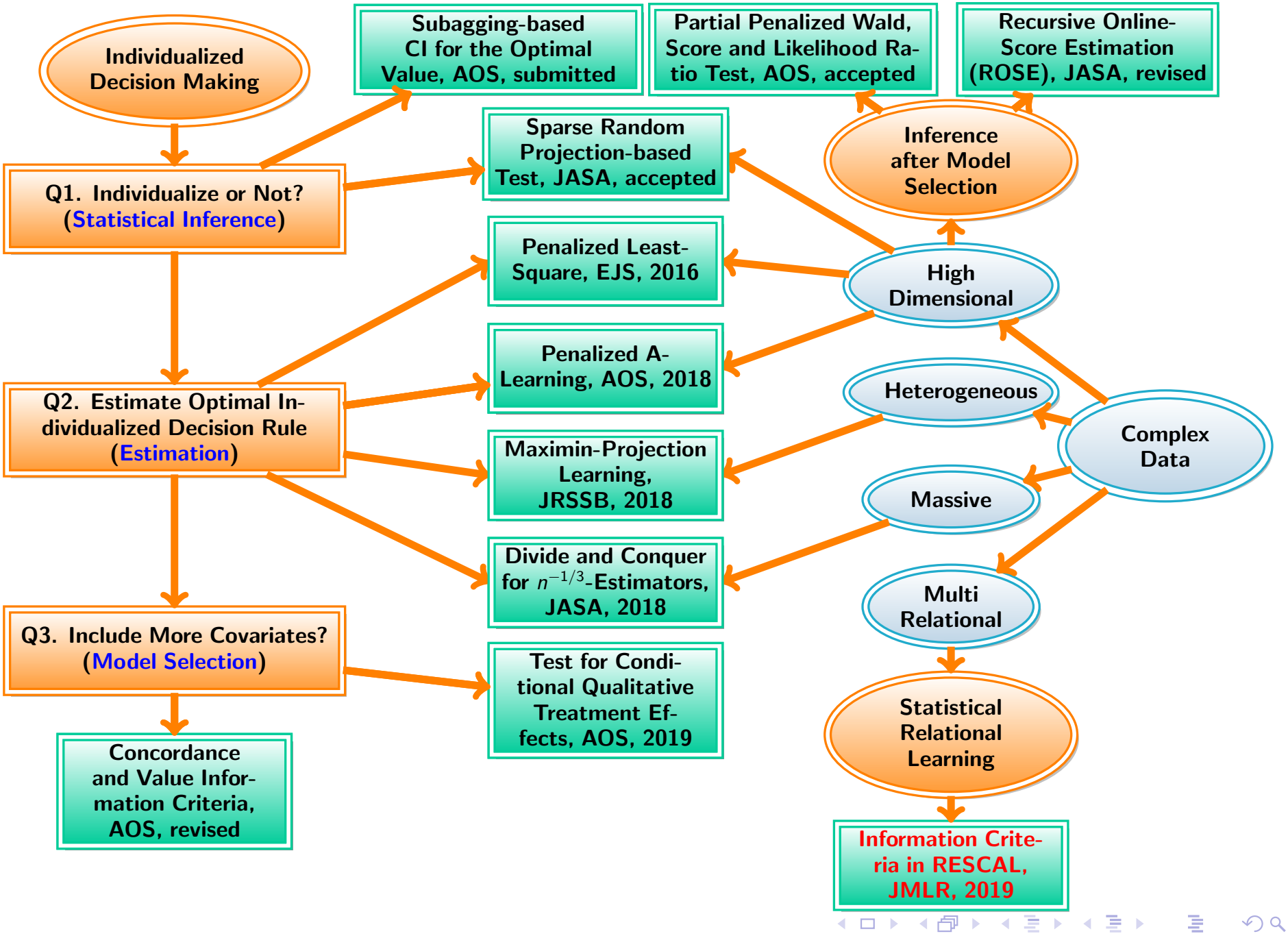


Our proposed CI

- achieves nominal coverage
- is on average shorter than that CI based on the one-step online method (Luedtke and van der Laan, 2016) and the m-out-of-n Bootstrap method (Chakraborty, B., Laber, E. B. and Zhao, 2014)
- is narrower than the CI based on the "oracle" method which works as if the optimal individualized decision rule were known


Individualized Decision Making
Q1. Individualize or Not? (Statistical Inference)
Q2. Estimate Optimal Individualized Decision Rule (Estimation)
Q3. Include More Covariates? (Model Selection)
Concordance and Value Information Criteria, AOS, revised
Subagging-based CI for the Optimal Value, AOS, submitted
Sparse Random Projection-based Test, JASA, accepted
Penalized Least-Square, EJS, 2016
Penalized A-Learning, AOS, 2018
Maximin-Projection Learning, JRSSB, 2018
Divide and Conquer for $n^{-1/3}$-Estimators, JASA, 2018
Test for Conditional Qualitative Treatment Effects, AOS, 2019
High Dimensional
Heterogeneous
Massive
Complex Data


Individualized Decision Making
Q1. Individualize or Not? (Statistical Inference)
Q2. Estimate Optimal Individualized Decision Rule (Estimation)
Q3. Include More Covariates? (Model Selection)
Concordance and Value Information Criteria, AOS, revised
Subagging-based CI for the Optimal Value, AOS, submitted
Sparse Random Projection-based Test, JASA, accepted
Penalized Least-Square, EJS, 2016
Penalized A-Learning, AOS, 2018
Maximin-Projection Learning, JRSSB, 2018
Divide and Conquer for $n^{-1/3}$-Estimators, JASA, 2018
Test for Conditional Qualitative Treatment Effects, AOS, 2019
Partial Penalized Wald, Score and Likelihood Ratio Test, AOS, accepted
Recursive Online-Score Estimation (ROSE), JASA, revised
Inference after Model Selection
High Dimensional
Heterogeneous
Massive
Complex Data


Individualized Decision Making
Q1. Individualize or Not? (Statistical Inference)
Q2. Estimate Optimal Individualized Decision Rule (Estimation)
Q3. Include More Covariates? (Model Selection)
Concordance and Value Information Criteria, AOS, revised
Subagging-based CI for the Optimal Value, AOS, submitted
Sparse Random Projection-based Test, JASA, accepted
Penalized Least-Square, EJS, 2016
Penalized A-Learning, AOS, 2018
Maximin-Projection Learning, JRSSB, 2018
Divide and Conquer for $n^{-1/3}$-Estimators, JASA, 2018
Test for Conditional Qualitative Treatment Effects, AOS, 2019
Partial Penalized Wald, Score and Likelihood Ratio Test, AOS, accepted
Recursive Online-Score Estimation (ROSE), JASA, revised
Inference after Model Selection
High Dimensional
Heterogeneous
Massive
Complex Data
Multi Relational
Statistical Relational Learning
Information Criteria in RESCAL, JMLR, 2019

Softwares

- **ITRSelect**: Variable Selection for Optimal Individualized Dynamic Treatment Regime (version 1.0-1). Available on CRAN.
- **ITRLearn**: Statistical Learning for Individualized Treatment Regime (version 1.0-1). Available on CRAN.
- **BayesSAE**: Bayesian Analysis of Small Area Estimation (version 1.0-2). Available on CRAN.
- **simplexreg**: Regression Analysis of Proportional Data Using Simplex Distributions (version 1.3). Available on CRAN.

Thank you!

Q-learning

Posit some model $h(x, \eta_g)$ for the baseline.

$$\text{Solve} \quad \sum_j \{\partial h(X_{gj}, \eta_g)/\partial \eta_g\}\{Y_{gj} - h_g(X_{gj}, \eta_g) - A_{gj}(X_{gj}^T \beta_g + c_0)\} = 0,$$

$$\sum_j A_{gj} X_{gj}\{Y_{gj} - h(X_{gj}, \eta_g) - A_{gj}(X_{gj}^T \beta_g + c_0)\} = 0,$$

$$\sum_g \sum_j A_{gj}\{Y_{gj} - h(X_{gj}, \eta_g) - A_{gj}(X_{gj}^T \beta_g + c_0)\} = 0.$$

A-learning

Posit some model $\pi(X, \alpha_g)$ for the propensity score, $h(x, \eta_g)$ for the baseline.

$$\text{Solve} \quad \sum [\pi(X, \alpha_g)\{1 - \pi(X, \alpha_g)\}]^{-1}\{\partial \pi(X, \alpha_g)/\partial \alpha_g\}\{A_{gj} - \pi(X, \alpha_g)\} = 0,$$

$$\sum_j \{\partial h(X_{gj}, \eta_g)/\partial \eta_g\}\{Y_{gj} - h(X_{gj}, \eta_g) - A_{gj}(X_{gj}^T \beta_g + c_0)\} = 0,$$

$$\sum_j X_{gj}\{A_{gj} - \pi(X_{gj}, \alpha_g)\}\{Y_{gj} - h(X_{gj}, \eta_g) - A_{gj}(X_{gj}^T \beta_g + c_0)\} = 0,$$

$$\sum_g \sum_j \{A_{gj} - \pi(X_{gj}, \alpha_g)\}\{Y_{gj} - h(X_{gj}, \eta_g) - A_{gj}(X_{gj}^T \beta_g + c_0)\} = 0.$$

Additional simulation results

Table: **Biases**, **standard deviations** (in parenthesis) of $\hat{\theta}^M$ and **coverage probabilities** (CP) of 95% Wald-type confidence intervals for θ^M.

Sce I	$\hat{\theta}_1^M$	$\hat{\theta}_2^M$	$\hat{\theta}_3^M$	CP(θ_1^M)	CP(θ_2^M)	CP(θ_3^M)
S1	$-0.001(0.011)$	$0.001(0.026)$	$0.0004(0.045)$	95.8%	96.2%	95.3%
S2	$-0.002(0.022)$	$0.001(0.052)$	$-0.002(0.083)$	93.5%	94.0%	96.0%
S3	$-0.002(0.018)$	$0.001(0.042)$	$-0.001(0.065)$	95.7%	95.8%	96.2%
S4	$-0.004(0.032)$	$0.0001(0.078)$	$-0.001(0.115)$	93.2%	95.0%	98.5%
Sce II	$\hat{\theta}_1^M$	$\hat{\theta}_2^M$	$\hat{\theta}_3^M$	CP(θ_1^M)	CP(θ_2^M)	CP(θ_3^M)
S1	$-0.002(0.036)$	$0.0002(0.036)$	$0.0002(0.023)$	95.5%	95.5%	95.3%
S2	$-0.009(0.061)$	$0.003(0.060)$	$-0.001(0.043)$	96.0%	96.0%	93.8%
S3	$-0.010(0.091)$	$-0.002(0.089)$	$-0.001(0.033)$	93.7%	93.7%	94.5%
S4	$-0.029(0.136)$	$0.034(0.130)$	$-0.002(0.056)$	98.3%	98.3%	95.0%

Health assessment questionnaire (HAQ) progression data

- An observational study to investigate the influence of early disease modifying antirheumatic drug (DMARD) treatment for patients with recent onset inflammatory polyarthritis (Farragher et al., 2010).
- 847 patients enrolled from 1990 to 2000.
- Treatments: methotrexate combination ($A = 1$) v.s. methotrexate monotherapy ($A = 0$).
- Response: reduction in HAQ scores between baseline and 5 years.
- Covariates: number of swollen joints, number of tender joints.
- Patients enrolled at different times showing heterogeneity; we considered three groups: 1990 - 1992; 1993 - 1996; 1997 - 2000.
- Hydroxychloroquine was increasingly used for the methotrexate combination treatment in the UK.

Health Assessment Quanstionnaire Data

Table: ITRs based on maximin-projection learning and random effects meta-analyses, and their estimated value functions.

Testing group	Group 1		Group 2		Group 3	
	maximin	**random**	**maximin**	**random**	**maximin**	**random**
$\hat{\theta}_1$	-0.48	0.00	0.61	0.16	-0.02	-0.01
$\hat{\theta}_2$	0.88	0.23	0.79	0.14	1.00	0.10
$\hat{\theta}_3$	-0.87	-0.12	-2.38	-0.11	-3.08	-0.32
$\hat{\mathrm{E}}Y_g^\star(d)$	-0.08	-0.09	**-0.05**	**-0.13**	-0.25	-0.25