Maximin-Projection Learning for Optimal Treatment Decision with Heterogeneous Individualized Treatment Effects

Chengchun Shi & Rui Song & Wenbin Lu & Bo Fu

North Carolina State University & Fudan University

Introduction

There are G groups of patients with

$$Y_g = h_g(X_g) + A_g(X_g^T\beta_g + c_0) + e_g.$$

Subgroup heterogeneity:

- ✦ Baseline function $h_g(\cdot)$.
- ✦ Propensity score function $\pi_g(\cdot)$.
- ✦ Individualized treatment effects β_g.
- ✦ Distribution of the error e_g.

Groupwise optimal treatment decision:

$$\mathbb{I}(x^T\beta_g + c_0 > 0).$$

Objective: Find a single treatment decision rule works reliably for future patients.

Motivation

Health assessment questionaire progression data:

- ✦ Methotrexate combinations ($A_g = 1$) v.s methotrexate monotherapy ($A_g = 0$)
- ✦ Group 1: patients enrolled from 1990 to 1992
- ✦ Group 2: patients enrolled from 1993 to 1996
- ✦ Group 3: patients enrolled from 1997 to 2000
- ✦ **Heterogeneity due to patients' enrollment time**

	Group 1	Group 2	Group 3
$\beta_g^{(1)}$	0.05(0.11)	-0.40(0.17)	0.07(0.21)
$\beta_g^{(2)}$	0.07(0.11)	0.06(0.21)	0.32(0.16)

The schizophrenia data:

- ✦ Cognitive-behavioural therapy ($A_g = 1$) v.s supportive counselling ($A_g = 0$)
- ✦ Group 1: patients from Manchester
- ✦ Group 2: patients from Liverpool
- ✦ Group 3: patients from North Nottinghamshire
- ✦ **Heterogeneity due to patients' treatment centre**

	Group 1	Group 2	Group 3
$\beta_g^{(1)}$	1.35(10.21)	1.17(11.52)	-20.73(13.27)
$\beta_g^{(2)}$	7.87(10.39)	-10.56(8.84)	3.45(9.14)

Methods

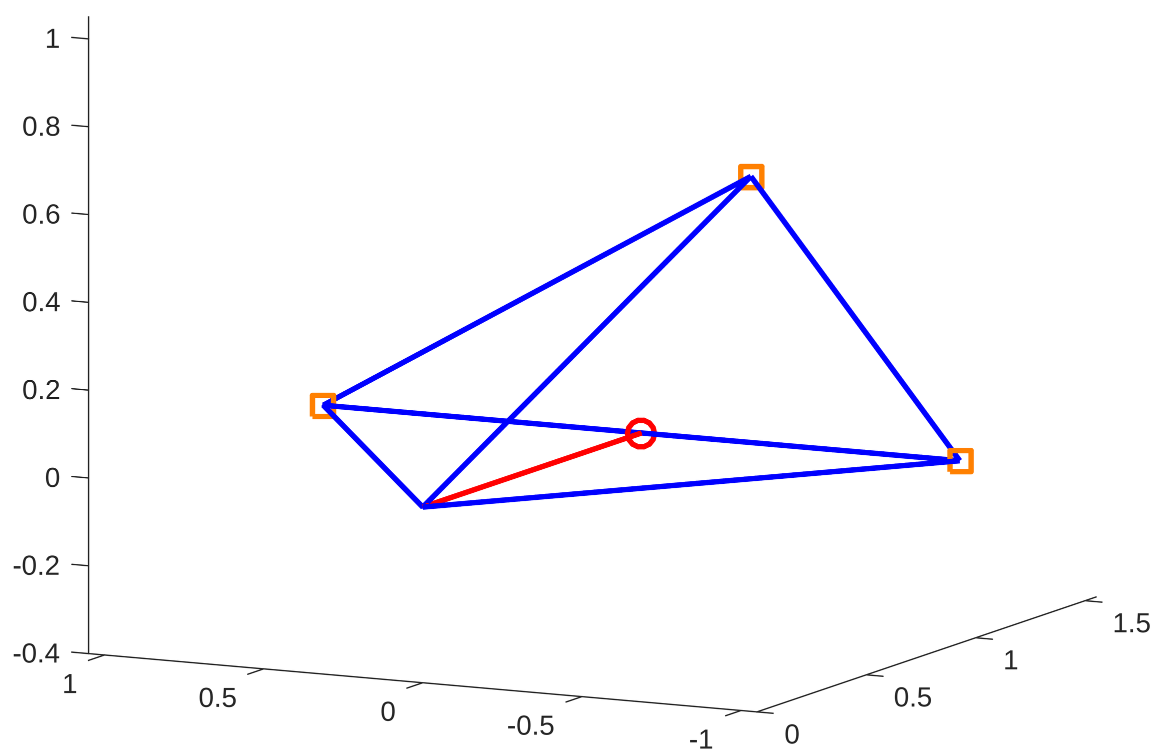
✦ **Random effects meta-analyses** (DerSimonian and Laird, 1986) estimate

$$\beta^R = \frac{1}{G}\sum_{g=1}^{G}\beta_g.$$

✦ **Consider the maximin effects:**

$$\beta^M = \arg\max_{\|\beta\|_2\le 1}\min_{g\in\{1,\dots,G\}}\beta^T\beta_g.$$

Geometrically, β^M corresponds to the **optimal equicorrelated point** of a subset of $\{\beta_1,\dots,\beta_G\}$.



✦ **Maximin-projection treatment decision:**

$$\mathbb{I}(x^T\beta^M + c^M > 0),$$

where $c^M = c_0/\max_{\|\beta\|_2\le 1}\min_g \beta^T\beta_g$.

✦ **Observed datasets:**

$$\{(X_{gj}, A_{gj}, Y_{gj}), g = 1,\dots,G, j = 1,\dots,m_g\}.$$

✦ **Parameter estimation:**

✧ **Step 1:** Posit some parametric model $\pi_g(\cdot,\alpha_g)$ for the $\pi_g(\cdot)$ and $h_g(\cdot,\eta_g)$ for $h_g(\cdot)$. Estimate parameters by jointly solving:

$$\sum_{j=1}^{m_g}\frac{\partial\pi_g(X_{gj},\alpha_g)}{\partial\alpha_g}\frac{\{A_{gj}-\pi_g(X_{gj},\alpha_g)\}}{\pi_g(X_{gj},\alpha_g)\{1-\pi_g(X_{gj},\alpha_g)\}} = 0,$$

$$\sum_{j=1}^{m_g}\frac{\partial h_g(X_{gj},\eta_g)}{\partial\eta_g}\varepsilon_{gj}(\eta_g,\beta_g,c_0) = 0,$$

$$\sum_{j=1}^{m_g}X_{gj}\{A_{gj}-\pi_g(X_{gj},\alpha_g)\}\varepsilon_{gj}(\eta_g,\beta_g,c_0) = 0,$$

$$\sum_{j=1}^{m_g}\sum_{g=1}^{G}\{A_{gj}-\pi_g(X_{gj},\alpha_g)\}\varepsilon_{gj}(\eta_g,\beta_g,c_0) = 0.$$

where

$$\varepsilon_{gj}(\eta_g,\beta_g,c_0) = \{Y_{gj} - h_g(X_{gj},\eta_g) - A_g(X_{gj}^T\beta_g + c_0)\}.$$

✧ **Step 2:** Estimate β^M and c^M by

$$\hat\beta^M = \arg\max_{\|\beta\|_2\le 1}\min_{g\in\{1,\dots,G\}}\beta^T\hat\beta_g,$$

$$\hat c^M = \hat c_0/\min_g \hat\beta_g^T\hat\beta^M.$$

Step 2 is a QCLP (Lee et al., 2016) and can be efficiently computed. Output

$$d_M(x) = \mathbb{I}(x^T\hat\beta^M + \hat c^M > 0).$$

✧ **Step 3:** Inference for β^M and c^M via bootstrap.

1. Independently generate B bootstrap samples within each group,
$$\{(X_{gj}^{(b)}, A_{gj}^{(b)}, Y_{gj}^{(b)}), g = 1,\dots,G, j = 1,\dots,m_g\}, \quad (1)$$
for $b = 1,\dots,B$.
2. Calculate $\hat\beta^{M(b)}$ and $\hat c^{M(b)}$ based on (1).
3. Obtain confidence intervals based on quantiles of $(\hat\beta^{M(1)},\dots,\hat\beta^{M(B)})$ and $(\hat c^{M(1)},\dots,\hat c^{M(B)})$.

Statistical interpretation

Two optimality criteria:

✦ Value difference function:

$$\mathrm{VD}_g(\beta,c) = \mathrm{E}\{X_g^T\beta_g + c_0\}\mathbb{I}(X_g^T\beta > -c).$$

✦ Percentage of making correct decisions:

$$\mathrm{PCD}_g(\beta,c) = 1 - \mathrm{E}|\mathbb{I}(X_g^T\beta > -c) - d_g^{opt}(X_g)|,$$

where $d_g^{opt}(x) = \mathbb{I}(x^T\beta_g > -c_0)$.

Theorem 1: Under certain conditions,

$$(\beta^M, c^M) = \arg\max_{\beta,c}\min_{g\in\{1,\dots,G\}}\mathrm{VD}_g(\beta,c),$$

$$(\beta^M, c^M) = \arg\max_{\beta,c}\min_{g\in\{1,\dots,G\}}\mathrm{PCD}_g(\beta,c).$$

Statistical property

Theorem 2: Under certain conditions, we have

$$\hat\beta^M \overset{P}{\to} \beta^M \text{ and } \hat c^M \overset{P}{\to} c^M.$$

In addition, $\sqrt{n}(\hat\beta^M - \beta^M)$ and $\sqrt{n}(\hat c^M - c^M)$ are jointly asymptotically normal with mean zero and some covariance matrix V^M.

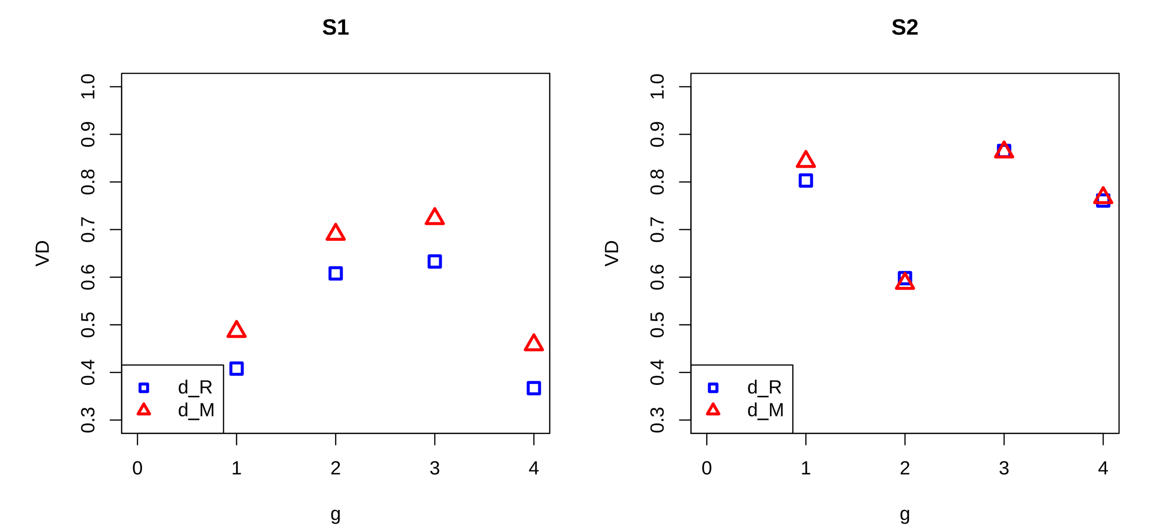
Simulations

✦ **Two scenarios** with different β_g's, $G = 4$.

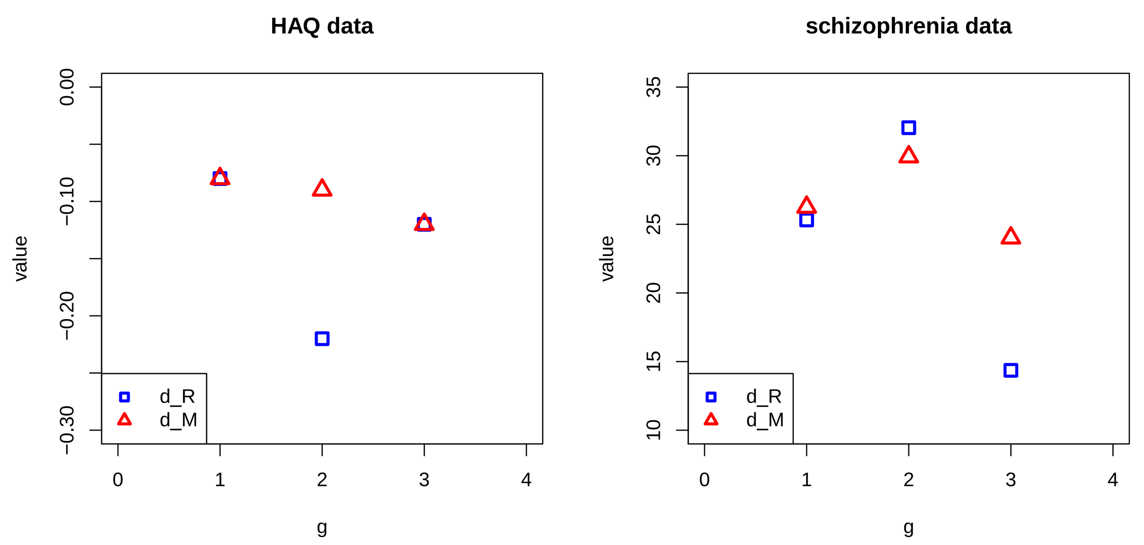
✧ **Scenario 1**: $\beta_1 = (2,0)^T$, $\beta_2 = 2(\cos(15°),\sin(15°))^T$, $\beta_3 = 2(\cos(70°),\sin(70°))^T$, $\beta_4 = (0,2)^T$.

✧ **Scenario 2**: $\beta_1 = 2.2(\cos(30°),\sin(30°))^T$, $\beta_2 = 1.5(\cos(45°),\sin(45°))^T$, $\beta_3 = 2.2(\cos(54°),\sin(54°))$, $\beta_4 = 2(\cos(60°),\sin(60°))$.

✦ Compare $d_M(\cdot)$ with the treatment regime based on random effects meta-analyses $d_R(\cdot)$.



Real data application



Contributions

- ✦ Propose a maximin-projection learning to aggregate optimal treatment decisions for patients from different populations with heterogeneity.
- ✦ Show the proposed treatment decision has nice statistical interpretation in the sense of maximizing the minimum PCD and value difference function.
- ✦ Provide a geometrical characterization of the maximin estimator.
- ✦ Study the statistical properties of the estimators.

NC STATE UNIVERSITY

cshi4@ncsu.edu