Supplement to “Maximin-Projection Learning for Optimal Treatment Decision with Heterogeneous Individualized Treatment Effects”

Chengchun Shi, Rui Song and Wenbin Lu

North Carolina State University, Raleigh, USA.

Bo Fu

Fudan University, Shanghai, People’s Republic of China.

In this supplement, we present a detailed discussion on the equicorrelated points set and the optimal equicorrelated point, additional numerical studies and proofs of theorem 1, theorem 3, theorem 4, lemma 3, theorem 5, and theorem 6.

B. More on the equicorrelated points set and the optimal equicorrelated point

For an arbitrary $s \times G$ matrix Ψ, let Ψ_g be its gth column vector. The equicorrelated points set of $\{\Psi_1, \dots, \Psi_G\}$ is defined by

$$\mathrm{E}(\Psi) = \left\{ t \in \mathbb{R}^s | t^T \Psi_j = t^T \Psi_i, \forall 1 \le i, j \le G \right\}.$$

To better understand $\mathrm{E}(\Psi)$, in Figure 1, we take $s = G = 2$, plot Ψ_1 and Ψ_2 as well as the triangle formed by these two vectors. We further plot the height of the triangle Ψ_0. Note that $\Psi_1^T \Psi_0 = \Psi_2^T \Psi_0$. In this small example, we have

$$\mathrm{E}(\Psi) = \left\{ a_0 \Psi_0 : a_0 \in \mathbb{R} \right\}.$$

Further assume $\|\Psi_1\|_2 = \|\Psi_2\|_2$. Then $\mathrm{E}(\Psi)$ consists of vectors that are parallel to the bisector of the angle formed by Ψ_1 and Ψ_2.

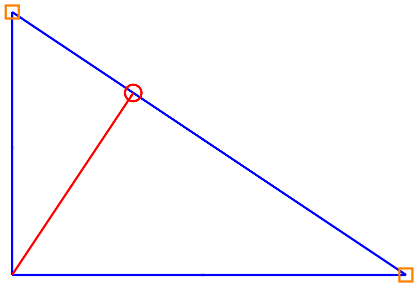
More generally, for any $t \in \mathrm{E}(\Psi)$, it follows from the definition of $\mathrm{E}(\Psi)$ that there exists some $\rho \in \mathbb{R}$,

$$\Psi^T t = \rho e, \tag{1}$$

where

$$e = (\underbrace{1, 1, \dots, 1}_{G}).$$

Fig. 1: Plots of Ψ_1 and Ψ_2 (denoted by the square symbol), and Ψ_0 (denoted by the circle symbol)



Obviously, we have $0_s \in \mathrm{E}(\Psi)$ where 0_s refers to an s-dimensional zero vector. Besides, for any $t_1, t_2 \in \mathrm{E}(\Psi)$, we have

$$\Psi^T t_1 = \rho_1 e \ \text{ and } \ \Psi^T t_2 = \rho_2 e,$$

for some constants $\rho_1, \rho_2 > 0$. Therefore,

$$\Psi^T(a_1 t_1 + a_2 t_2) = (a_1\rho_1 + a_2\rho_2)e,$$

for any $a_1, a_2 \in \mathbb{R}$. This implies $a_1 t_1 + a_2 t_2 \in \mathrm{E}(\Psi)$. Hence, $\mathrm{E}(\Psi)$ forms a linear subspace in $\mathbb{R}^s$.

Moreover, for any $t \in \mathrm{E}(\Psi)$, we can represented t by $t = t_0 + t_*$ for $t_0 \in C(\Psi)$ and $t_* \in N(\Psi^T)$ where $C(\Psi)$ denotes the column space of Ψ and $N(\Psi^T)$ denotes the null space of Ψ^T. By definition, we have $t_0 = \Psi\omega_0$ for some $\omega_0 \in \mathbb{R}^s$. It follows from (1) that

$$\Psi^T\Psi\omega_0 = \rho e.$$

Assume $e \in C(\Psi^T)$. Using some standard arguments in linear least square regressions, we have

$$\mathrm{E}(\Psi) = \left\{\rho\Psi[\Psi^T\Psi]^+ e + t_* : \rho \in \mathbb{R}, t_* \in N(\Psi^T)\right\}, \tag{2}$$

where $[\Psi^T\Psi]^+$ is the Moore-Penrose inverse of $\Psi^T\Psi$.

The optimal equicorrelated point is defined by

$$\mathrm{E}^\star(\Psi) = \arg\max_{\substack{t\in E(\Psi)\\ \|t\|_2=1}} \left\{t^T\Psi_g, \forall 1 \le g \le G\right\}.$$

By (2), finding $\mathrm{E}^\star(\Psi)$ is equivalent to the following problem

$$\arg\max_{\rho,t_*} \rho^2 e^T[\Psi^T\Psi]^+ e, \qquad s.t. \|t_*\|_2^2 + \rho^2 e^T[\Psi^T\Psi]^+ e = 1. \tag{3}$$

When $e^T[\Psi^T\Psi]^+ e > 0$, (3) is maximized at $t_* = 0$, $\rho = (e^T[\Psi^T\Psi]^+ e)^{-1/2}$. Therefore,

$$\mathrm{E}^\star(\Psi) = (e^T[\Psi^T\Psi]^+ e)^{-1/2}\Psi[\Psi^T\Psi]^+ e.$$

This proves lemma 2.

In the following, we show $e^T[\Psi^T\Psi]^+e > 0$. When $e \in C(\Psi^T)$, there exists some vector ω_* such that $e = \Psi^T\omega_*$. Hence, it is equivalent to show $\omega_*^T\Psi[\Psi^T\Psi]^+\Psi^T\omega_* > 0$. Note that $\Psi[\Psi^T\Psi]^+\Psi^T$ is the projection matrix of Ψ. If $\omega_*^T\Psi[\Psi^T\Psi]^+\Psi^T\omega_* = 0$, then ω_* belongs to the null space of Ψ^T. However, this implies $\Psi^T\omega_* = 0_G$ and we've reached a contradiction. Therefore, we have $e^T[\Psi^T\Psi]^+e > 0$.

C. Additional simulation results

C.1. Homogeneous individualized treatment effects

As suggested by one of the referee, we further examine our methods under settings where some of the β_g's are the same. As in Section 5, responses were generated from

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

for $j = 1, \dots, 200$, $g = 1, \dots, 4$, $X_{gj} = (X_{gj}^{(1)}, X_{gj}^{(2)})^T \overset{iid}{\sim} N(0, I_2)$ and $\varepsilon_{gj} \overset{iid}{\sim} N(0, 0.25)$. We consider two scenarios. In the first scenario, we set $\beta_1 = \beta_2 = (2, 0)^T$, and $\beta_3 = \beta_4 = (0, 2)^T$. In the second scenario, we set $\beta_1 = \beta_2 = \beta_3 = \beta_4 = (2\cos(45°), 2\sin(45°))^T$. By definition, we have

$$\beta_{(0)}^M = \arg\max_{\beta:\|\beta\|_2\le 1}\min_{g\in\{1,2,3,4\}} \beta^T\beta_g = \{\cos(45°), \sin(45°)\}^T,$$

and $c_{(0)}^M = c_0/(\beta_0^T\beta_{(0)}^M) = 0$.

We consider the same four settings as in Section 5. In Table 7, we report the biases and standard deviations of $\hat\beta^M$ and $\hat c^M$, based on 600 simulations. Confidence intervals are omitted since it remains unknown whether our bootstrap procedure is consistent under settings where some of the β_g's are the same. From Table 7, it is evident that $\hat\beta^M$ and $\hat c^M$ are consistent to $\beta_{(0)}^M$ and $c_{(0)}^M$ in both two scenarios, respectively.

We further evaluate the PCD and the VD under the estimated maximin OTR $\hat d_M(x) = I(x^T\hat\beta^M > -\hat c^M)$ and compare them with those under the estimated pooled OTR $\hat d_P(x) = I(x^T\hat\beta^P > -\hat c^P)$. The PCD and the VD under the OTR obtained by random effects meta-analyses are very close to those under the pooled OTR, and are hence omitted for brevity. Simulation results are summarized in Table 8, 9, 10 and 11. In Scenario 1, the estimated maximin OTR performs uniformly better than the estimated pooled OTR. The VD under $\hat d_M$ are approximately half larger than those under $\hat d_P$. In Scenario 2, since $\beta_g's$ are the same, we can show that $\hat\beta^P \xrightarrow{P} (2\cos(45°), 2\sin(45°))^T$ and $\hat c^P \xrightarrow{P} 0$ when either the propensity score model or the baseline mode is correct. The VD under the estimated maximin OTR are very similar to those under the estimated pooled OTR. Besides, they are very close to $\sqrt{2/\pi} \approx 0.798$, which corresponds to the VD under the groupwise OTR. The PCD under the estimated maximin OTR are sightly lower than those under the estimated pooled OTR. Nonetheless, they are above 96% for all cases. This implies the estimated maximin OTR is consistent to the groupwise OTR in a homogeneous setting.

Table 7: Biases, standard deviations (SD) of $\hat{\beta}^M$ and $\hat{c}^M$

		$\hat{\beta}_1^M$		$\hat{\beta}_2^M$		$\hat{c}^M$	
		Bias	SD	Bias	SD	Bias	SD
Scenario 1	Setting 1	-0.001	0.022	4.0×10^{-4}	0.022	3.0×10^{-4}	0.025
	Setting 2	-8.0×10^{-4}	0.043	-0.002	0.043	-0.001	0.048
	Setting 3	-0.003	0.030	-0.002	0.029	0.001	0.037
	Setting 4	-0.003	0.059	-0.002	0.059	-0.002	0.065
Scenario 2	Setting 1	-0.001	0.026	-1.0×10^{-4}	0.026	2.0×10^{-4}	0.018
	Setting 2	-0.002	0.045	-4.0×10^{-4}	0.045	-0.001	0.034
	Setting 3	-0.003	0.057	-0.002	0.057	-4.0×10^{-4}	0.026
	Setting 4	-0.004	0.094	-0.009	0.095	-0.001	0.046

Table 8: VD results (with standard errors in parenthesis) for Scenario 1 under the maximin and pooled OTRs.

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	0.355(0.002)	0.355(0.002)	0.356(0.002)	0.357(0.002)
	maximin	0.556(0.001)	0.554(0.001)	0.555(0.001)	0.554(0.001)
Setting 2	pooled	0.354(0.003)	0.355(0.003)	0.355(0.003)	0.355(0.003)
	maximin	0.546(0.002)	0.547(0.002)	0.544(0.002)	0.545(0.002)
Setting 3	pooled	0.358(0.004)	0.355(0.003)	0.355(0.003)	0.355(0.004)
	maximin	0.552(0.001)	0.550(0.001)	0.553(0.001)	0.550(0.001)
Setting 4	pooled	0.357(0.004)	0.355(0.004)	0.356(0.004)	0.353(0.004)
	maximin	0.536(0.002)	0.539(0.002)	0.537(0.002)	0.536(0.002)

Table 9: PCD results (%, with standard errors in parenthesis) for Scenario 1 under the maximin and pooled OTRs.

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	64.7(0.1)	64.7(0.1)	64.8(0.1)	64.8(0.1)
	maximin	74.6(<0.1)	74.5(<0.1)	74.5(<0.1)	74.5(<0.1)
Setting 2	pooled	64.7(0.1)	64.7(0.1)	64.7(0.1)	64.8(0.1)
	maximin	74.1(0.1)	74.1(0.1)	74.0(0.1)	74.0(0.1)
Setting 3	pooled	64.9(0.2)	64.8(0.2)	64.8(0.2)	64.8(0.2)
	maximin	74.4(0.1)	74.2(0.1)	74.4(0.1)	74.3(0.1)
Setting 4	pooled	65.0(0.2)	64.9(0.2)	64.9(0.2)	64.8(0.2)
	maximin	73.6(0.1)	73.8(0.1)	73.6(0.1)	73.6(0.1)

Table 10: VD results (with standard errors in parenthesis) for Scenario 2 under the maximin and pooled OTRs.

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	0.798(<0.001)	0.798(<0.001)	0.797(<0.001)	0.798(<0.001)
	maximin	0.797(<0.001)	0.797(<0.001)	0.797(<0.001)	0.797(<0.001)
Setting 2	pooled	0.797(<0.001)	0.797(<0.001)	0.797(<0.001)	0.797(<0.001)
	maximin	0.796(<0.001)	0.796(<0.001)	0.796(<0.001)	0.796(<0.001)
Setting 3	pooled	0.797(<0.001)	0.797(<0.001)	0.797(<0.001)	0.797(<0.001)
	maximin	0.795(<0.001)	0.795(<0.001)	0.795(<0.001)	0.795(<0.001)
Setting 4	pooled	0.794(<0.001)	0.794(<0.001)	0.794(<0.001)	0.794(<0.001)
	maximin	0.790(<0.001)	0.790(<0.001)	0.790(<0.001)	0.789(<0.001)

Table 11: PCD results (%, with standard errors in parenthesis) for Scenario 2 under the maximin and pooled OTRs.

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	99.1(<0.1)	99.1(<0.1)	99.1(<0.1)	99.1(<0.1)
	maximin	98.8(<0.1)	98.8(<0.1)	98.8(<0.1)	98.8(<0.1)
Setting 2	pooled	98.6(<0.1)	98.5(<0.1)	98.5(<0.1)	98.5(<0.1)
	maximin	98.0(<0.1)	97.9(<0.1)	98.0(<0.1)	98.0(<0.1)
Setting 3	pooled	98.3(<0.1)	98.3(<0.1)	98.4(<0.1)	98.3(<0.1)
	maximin	97.7(0.1)	97.7(0.1)	97.7(0.1)	97.7(0.1)
Setting 4	pooled	97.4(0.1)	97.2(0.1)	97.3(0.1)	97.1(0.1)
	maximin	96.1(0.1)	96.2(0.1)	96.2(0.1)	96.0(0.1)

C.2. Non-normal covariates

We further examine the robustness of our estimator with non-normal covariates. We consider the same model as in Section 5,

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

for $j = 1, \dots, 200$, $g = 1, \dots, 4$, where $\varepsilon_{gj} \overset{iid}{\sim} N(0, 0.25)$. Covariates $X_{gj} = (X_{gj}^{(1)}, X_{gj}^{(2)})^T$ were generated from the following distributions:

(i) $X_{gj}^{(k)} \overset{iid}{\sim} t(4)/\sqrt{2}$, for $j = 1, \dots, 200, g = 1, \dots, 4, k = 1, 2$, where $t(k)$ stands for the Student's t-distribution with k degrees of freedoms.

(ii) $X_{gj}^{(1)} \overset{iid}{\sim} 2\text{Ber}(0.5) - 1$, $X_{gj}^{(2)} \overset{iid}{\sim} N(0, 1)$, for $j = 1, \dots, 200, g = 1, \dots, 4$, where $\text{Ber}(p)$ denotes the Bernoulli random variable with success probability p. Besides, $X_{g_1 j_1}^{(1)}$ and $X_{g_2 j_2}^{(2)}$ are independent, for any g_1, g_2, j_1, j_2.

(iii) $X_{1j}^{(k)} \overset{iid}{\sim} N(0, 1)$, $X_{2j}^{(k)} \overset{iid}{\sim} t(3)/\sqrt{3}$, $X_{3j}^{(k)} \overset{iid}{\sim} t(4)/\sqrt{2}$, $X_{4j}^{(k)} \overset{iid}{\sim} t(5)/\sqrt{5/3}$, for $j = 1, \dots, 200$, $g = 1, \dots, 4$, $k = 1, 2$. Besides, X_{1j_1}, X_{2j_2}, X_{3j_3} and X_{4j_4} are independent for any j_1, j_2, j_3, j_4.

Note that in (iii), the distributions of the covariates are allowed to vary across different groups. We consider the same two scenarios for β_g's, and the same four settings for the propensity score models and the baseline models as in Section 5. We conduct 600 simulation replications. In Table 12, 13 and 14, we report the biases and standard deviations of $\hat{\beta}^M$ and $\hat{c}^M$, as well as the coverage probabilities (CP) of 95% Wald-type confidence intervals for $\beta_{(0)}^M$ and $c_{(0)}^M$ when covariates are generated as in (i), (ii) and (iii), respectively. The confidence intervals are calculated based on 600 bootstrap samples. Findings are similar to those with normal covariates.

Table 12: Biases, standard deviations (in parenthesis) of $\hat{\beta}^M$, $\hat{c}^M$ and coverage probabilities (CP) of 95% Wald-type confidence intervals for $\beta^M_{(0)}$, $c^M_{(0)}$ when covariates are generated as in (i).

Scenario 1	$\hat{\beta}^M_1$	$\hat{\beta}^M_2$	$\hat{c}^M$	CP for $\hat{\beta}^M_1$	CP for $\hat{\beta}^M_2$	CP for $\hat{c}^M$
Setting 1	-0.002(0.02)	0.001(0.019)	-0.0004(0.026)	97.7%	97.7%	94.2%
Setting 2	-0.002(0.035)	-0.0001(0.035)	-0.002(0.044)	96.2%	96.2%	94.5%
Setting 3	-0.003(0.035)	0.001(0.035)	-0.002(0.038)	95.2%	95.2%	94.3%
Setting 4	-0.003(0.06)	-0.002(0.061)	-0.006(0.071)	96.8%	96.8%	92.7%
Scenario 2	$\hat{\beta}^M_1$	$\hat{\beta}^M_2$	$\hat{c}^M$	CP for $\hat{\beta}^M_1$	CP for $\hat{\beta}^M_2$	CP for $\hat{c}^M$
Setting 1	-0.001(0.027)	-0.0003(0.027)	-0.0004(0.025)	97.2%	97.2%	94.2%
Setting 2	0.002(0.044)	-0.005(0.044)	-0.002(0.041)	96.2%	96.2%	94.5%
Setting 3	-0.008(0.086)	-0.003(0.085)	-0.002(0.036)	95.0%	95.0%	94.2%
Setting 4	-0.027(0.133)	0.003(0.127)	-0.005(0.065)	97.8%	97.8%	92.7%

Table 13: Biases, standard deviations (in parenthesis) of $\hat{\beta}^M$, $\hat{c}^M$ and coverage probabilities (CP) of 95% Wald-type confidence intervals for $\beta^M_{(0)}$, $c^M_{(0)}$ when covariates are generated as in (ii).

Scenario 1	$\hat{\beta}^M_1$	$\hat{\beta}^M_2$	$\hat{c}^M$	CP for $\hat{\beta}^M_1$	CP for $\hat{\beta}^M_2$	CP for $\hat{c}^M$
Setting 1	-0.0004(0.026)	-0.001(0.026)	-0.001(0.025)	96.7%	96.7%	93.8%
Setting 2	-0.00001(0.047)	-0.003(0.046)	-0.001(0.039)	94.8%	94.8%	97.3%
Setting 3	-0.001(0.032)	-0.001(0.032)	-0.0001(0.037)	94.7%	94.7%	94.8%
Setting 4	-0.003(0.055)	-0.001(0.053)	-0.023(0.046)	95.3%	95.3%	93.3%
Scenario 2	$\hat{\beta}^M_1$	$\hat{\beta}^M_2$	$\hat{c}^M$	CP for $\hat{\beta}^M_1$	CP for $\hat{\beta}^M_2$	CP for $\hat{c}^M$
Setting 1	-0.001(0.032)	-0.0002(0.032)	-0.001(0.024)	96.8%	96.8%	93.8%
Setting 2	-0.0003(0.061)	-0.005(0.062)	-0.001(0.037)	98.0%	98.0%	97.3%
Setting 3	-0.007(0.086)	-0.003(0.084)	-0.00003(0.034)	93.8%	93.8%	94.8%
Setting 4	-0.007(0.118)	-0.014(0.121)	-0.021(0.043)	96.8%	96.8%	93.3%

Table 14: Biases, standard deviations (in parenthesis) of $\hat{\beta}^M$, $\hat{c}^M$ and coverage probabilities (CP) of 95% Wald-type confidence intervals for $\beta^M_{(0)}$, $c^M_{(0)}$ when covariates are generated as in (iii).

Scenario 1	$\hat{\beta}^M_1$	$\hat{\beta}^M_2$	$\hat{c}^M$	CP for $\hat{\beta}^M_1$	CP for $\hat{\beta}^M_2$	CP for $\hat{c}^M$
Setting 1	0.0004(0.027)	-0.001(0.027)	0.001(0.025)	97.5%	97.5%	95.0%
Setting 2	0.001(0.05)	-0.005(0.05)	0.001(0.042)	97.0%	97.0%	94.0%
Setting 3	0.0002(0.038)	-0.002(0.038)	-0.001(0.034)	96.5%	96.5%	94.8%
Setting 4	-0.002(0.074)	-0.006(0.076)	-0.00005(0.051)	95.8%	95.8%	95.7%
Scenario 2	$\hat{\beta}^M_1$	$\hat{\beta}^M_2$	$\hat{c}^M$	CP for $\hat{\beta}^M_1$	CP for $\hat{\beta}^M_2$	CP for $\hat{c}^M$
Setting 1	0.001(0.042)	-0.003(0.042)	0.001(0.024)	97.3%	97.3%	95.0%
Setting 2	-0.006(0.084)	-0.004(0.083)	0.001(0.041)	97.7%	97.7%	94.0%
Setting 3	-0.004(0.088)	-0.007(0.088)	-0.001(0.032)	96.5%	96.5%	94.8%
Setting 4	-0.020(0.137)	-0.006(0.132)	0.0001(0.047)	98.0%	98.0%	95.7%

In Table 15-20, we present the VD under the estimated maximin OTR and the estimated pooled OTR. The PCD results are reported in Table 23-28 in Section C.3. The VD and the PCD under the OTR estimated by random effects meta-analyses are omitted, since they are very close to those under the pooled OTR. It can be seen that in Scenario 1, the maximin OTR are uniformly better than the pooled OTR over all groups. In Scenario 2, the VD under the maximin OTR are larger than those under the pooled OTR when the first group is taken as the testing group. When other groups are taken as the testing groups, the VD under the maximin and the pooled OTRs become comparable.

Table 15: The VD results (with standard errors in parenthesis) under the estimated maximin OTR and the pooled OTR for Scenario 1 when covariates are generated as in (i).

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	0.410(0.002)	0.579(0.002)	0.598(0.002)	0.371(0.002)
	maximin	0.465(0.001)	0.637(<0.001)	0.667(<0.001)	0.445(0.001)
Setting 2	pooled	0.409(0.002)	0.578(0.002)	0.595(0.002)	0.368(0.002)
	maximin	0.464(0.002)	0.636(0.001)	0.664(0.001)	0.439(0.002)
Setting 3	pooled	0.409(0.002)	0.575(0.002)	0.598(0.002)	0.377(0.002)
	maximin	0.464(0.001)	0.637(0.001)	0.665(0.001)	0.443(0.002)
Setting 4	pooled	0.406(0.003)	0.572(0.002)	0.595(0.002)	0.377(0.003)
	maximin	0.456(0.003)	0.633(0.001)	0.659(0.001)	0.427(0.003)

Table 16: The VD results (with standard errors in parenthesis) under the estimated maximin OTR and the pooled OTR for Scenario 2 when covariates are generated as in (i).

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	0.752(<0.001)	0.551(<0.001)	0.798(<0.001)	0.704(<0.001)
	maximin	0.782(<0.001)	0.545(<0.001)	0.798(<0.001)	0.709(<0.001)
Setting 2	pooled	0.751(<0.001)	0.551(<0.001)	0.797(<0.001)	0.703(<0.001)
	maximin	0.779(0.001)	0.544(<0.001)	0.796(<0.001)	0.706(0.001)
Setting 3	pooled	0.751(0.001)	0.551(<0.001)	0.797(<0.001)	0.704(<0.001)
	maximin	0.779(0.001)	0.545(<0.001)	0.795(0.001)	0.706(0.001)
Setting 4	pooled	0.748(0.001)	0.550(<0.001)	0.795(<0.001)	0.702(0.001)
	maximin	0.767(0.002)	0.543(<0.001)	0.789(0.001)	0.698(0.001)

Table 17: The VD results (with standard errors in parenthesis) under the estimated maximin OTR and the pooled OTR for Scenario 1 when covariates are generated as in (ii).

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	0.447(0.002)	0.657(0.002)	0.523(0.002)	0.132(0.002)
	maximin	0.559(0.002)	0.785(0.001)	0.686(0.001)	0.284(0.002)
Setting 2	pooled	0.447(0.002)	0.654(0.002)	0.525(0.003)	0.137(0.003)
	maximin	0.558(0.003)	0.781(0.002)	0.685(0.002)	0.286(0.004)
Setting 3	pooled	0.450(0.003)	0.658(0.003)	0.525(0.004)	0.139(0.004)
	maximin	0.558(0.002)	0.785(0.001)	0.685(0.001)	0.282(0.003)
Setting 4	pooled	0.451(0.004)	0.658(0.004)	0.525(0.004)	0.146(0.005)
	maximin	0.554(0.004)	0.778(0.002)	0.684(0.002)	0.287(0.004)

Table 18: The VD results (with standard errors in parenthesis) under the estimated maximin OTR and the pooled OTR for Scenario 2 when covariates are generated as in (ii).

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	0.846(0.001)	0.617(<0.001)	0.869(<0.001)	0.745(<0.001)
	maximin	0.915(0.001)	0.604(<0.001)	0.870(0.001)	0.758(0.001)
Setting 2	pooled	0.846(0.001)	0.616(<0.001)	0.868(0.001)	0.744(0.001)
	maximin	0.913(0.001)	0.604(0.001)	0.864(0.001)	0.751(0.001)
Setting 3	pooled	0.845(0.001)	0.616(<0.001)	0.866(0.001)	0.741(0.001)
	maximin	0.907(0.002)	0.606(<0.001)	0.862(0.001)	0.751(0.002)
Setting 4	pooled	0.844(0.002)	0.614(<0.001)	0.862(0.001)	0.737(0.002)
	maximin	0.900(0.003)	0.606(0.001)	0.847(0.002)	0.735(0.003)

Table 19: The VD results (with standard errors in parenthesis) under the estimated maximin OTR and the pooled OTR for Scenario 1 when covariates are generated as in (iii).

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	0.388(0.003)	0.537(0.001)	0.604(0.001)	0.378(0.002)
	maximin	0.483(0.001)	0.584(<0.001)	0.667(<0.001)	0.453(0.001)
Setting 2	pooled	0.388(0.003)	0.536(0.001)	0.602(0.001)	0.377(0.002)
	maximin	0.481(0.003)	0.583(0.001)	0.664(0.001)	0.445(0.002)
Setting 3	pooled	0.395(0.003)	0.537(0.002)	0.599(0.002)	0.371(0.002)
	maximin	0.481(0.002)	0.584(0.001)	0.666(0.001)	0.450(0.001)
Setting 4	pooled	0.397(0.004)	0.535(0.002)	0.594(0.002)	0.366(0.003)
	maximin	0.476(0.003)	0.581(0.001)	0.660(0.001)	0.435(0.003)

Table 20: The VD results (with standard errors in parenthesis) under the estimated maximin OTR and the pooled OTR for Scenario 2 when covariates are generated as in (iii).

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	0.801(<0.001)	0.506(<0.001)	0.798(<0.001)	0.724(<0.001)
	maximin	0.845(0.001)	0.500(<0.001)	0.799(<0.001)	0.731(<0.001)
Setting 2	pooled	0.800(0.001)	0.505(<0.001)	0.797(<0.001)	0.724(<0.001)
	maximin	0.838(0.001)	0.500(<0.001)	0.796(0.001)	0.728(0.001)
Setting 3	pooled	0.802(0.001)	0.505(<0.001)	0.797(<0.001)	0.723(<0.001)
	maximin	0.840(0.001)	0.501(<0.001)	0.795(0.001)	0.727(0.001)
Setting 4	pooled	0.801(0.001)	0.504(<0.001)	0.795(<0.001)	0.721(0.001)
	maximin	0.827(0.002)	0.499(<0.001)	0.788(0.001)	0.717(0.002)

C.3. Additional tables

Table 21: The PCD results (%, with standard errors in parenthesis) for Scenario 1 under the estimated maximin OTR $\hat{d}_M$, the pooled OTR $\hat{d}_P$ and the OTR estimated by random effects meta-analyses $\hat{d}_R$.

Testing group		First group	Second group	Third group	Fourth group
Setting 1	$\hat{d}_P$	67.1(0.1)	77.6(0.1)	79.2(0.1)	65.3(0.1)
	$\hat{d}_R$	67.1(<0.1)	77.6(<0.1)	79.2(<0.1)	65.2(<0.1)
	$\hat{d}_M$	70.9(0.1)	83.3(0.1)	86.1(0.1)	69.5(0.1)
Setting 2	$\hat{d}_P$	67.1(0.1)	77.6(0.1)	79.1(0.1)	65.2(0.1)
	$\hat{d}_R$	67.0(<0.1)	77.6(<0.1)	79.2(<0.1)	65.2(<0.1)
	$\hat{d}_M$	70.8(0.1)	83.3(0.1)	85.9(0.1)	69.3(0.1)
Setting 3	$\hat{d}_P$	67.1(0.1)	77.6(0.1)	79.2(0.1)	65.3(0.1)
	$\hat{d}_R$	67.0(0.1)	77.5(0.1)	79.2(0.1)	65.2(0.1)
	$\hat{d}_M$	70.8(0.1)	83.2(0.1)	86.2(0.1)	69.3(0.1)
Setting 4	$\hat{d}_P$	67.1(0.2)	77.6(0.2)	79.3(0.2)	65.3(0.2)
	$\hat{d}_R$	67.1(0.1)	77.5(0.1)	79.2(0.1)	65.2(0.1)
	$\hat{d}_M$	70.4(0.2)	83.2(0.1)	85.7(0.1)	68.7(0.2)

Table 22: The PCD results (%, with standard errors in parenthesis) for Scenario 2 under the estimated maximin OTR $\hat{d}_M$, the pooled OTR $\hat{d}_P$ and the OTR estimated by random effects meta-analyses $\hat{d}_R$.

Testing group		First group	Second group	Third group	Fourth group
Setting 1	$\hat{d}_P$	86.8(<0.1)	98.2(<0.1)	94.6(<0.1)	90.4(<0.1)
	$\hat{d}_R$	86.8(<0.1)	98.4(<0.1)	94.6(<0.1)	90.4(<0.1)
	$\hat{d}_M$	91.7(0.1)	94.6(0.1)	94.9(0.1)	91.6(0.1)
Setting 2	$\hat{d}_P$	86.8(0.1)	97.9(<0.1)	94.5(0.1)	90.3(0.1)
	$\hat{d}_R$	86.8(<0.1)	98.3(<0.1)	94.6(<0.1)	90.4(<0.1)
	$\hat{d}_M$	91.5(0.1)	94.6(0.1)	94.8(0.1)	91.5(0.1)
Setting 3	$\hat{d}_P$	86.8(0.1)	97.8(<0.1)	94.6(0.1)	90.4(0.1)
	$\hat{d}_R$	86.7(0.1)	97.8(<0.1)	94.6(0.1)	90.4(0.1)
	$\hat{d}_M$	91.5(0.1)	94.8(0.1)	94.5(0.1)	91.6(0.1)
Setting 4	$\hat{d}_P$	86.7(0.1)	96.9(0.1)	94.3(0.1)	90.3(0.1)
	$\hat{d}_R$	87.0(0.1)	97.8(<0.1)	94.4(0.1)	90.2(0.1)
	$\hat{d}_M$	90.3(0.2)	94.7(0.1)	93.7(0.2)	91.0(0.2)

Table 23: The PCD results (%, with standard errors in parenthesis) under the estimated maximin OTR and the pooled OTR for Scenario 1 when covariates are generated as in (i).

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	68.0(0.1)	79.4(0.1)	80.8(0.1)	65.9(0.1)
	maximin	71.3(0.1)	84.3(<0.1)	87.2(<0.1)	70.0(0.1)
Setting 2	pooled	68.0(0.1)	79.4(0.1)	80.6(0.1)	65.7(0.1)
	maximin	71.3(0.1)	84.3(0.1)	87.0(0.1)	69.7(0.1)
Setting 3	pooled	68.0(0.1)	79.2(0.1)	80.9(0.1)	66.2(0.1)
	maximin	71.3(0.1)	84.4(0.1)	87.0(0.1)	70.0(0.1)
Setting 4	pooled	67.9(0.1)	79.0(0.2)	80.7(0.2)	66.3(0.2)
	maximin	70.9(0.2)	84.2(0.1)	86.6(0.1)	69.2(0.2)

Table 24: The PCD results (%, with standard errors in parenthesis) under the estimated maximin OTR and the pooled OTR for Scenario 2 when covariates are generated as in (i).

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	88.3(<0.1)	98.3(<0.1)	95.1(<0.1)	91.3(<0.1)
	maximin	92.6(0.1)	95.2(0.1)	95.3(0.1)	92.3(0.1)
Setting 2	pooled	88.3(0.1)	97.8(<0.1)	94.9(0.1)	91.1(0.1)
	maximin	92.4(0.1)	95.0(0.1)	94.9(0.1)	92.0(0.1)
Setting 3	pooled	88.2(0.1)	97.8(<0.1)	95.0(0.1)	91.3(0.1)
	maximin	92.5(0.1)	95.4(0.1)	94.9(0.1)	92.1(0.1)
Setting 4	pooled	88.1(0.1)	97.1(0.1)	94.6(0.1)	91.3(0.1)
	maximin	91.4(0.2)	94.9(0.1)	94(0.2)	91.4(0.2)

Table 25: The PCD results (%, with standard errors in parenthesis) under the estimated maximin OTR and the pooled OTR for Scenario 1 when covariates are generated as in (ii).

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	72.3(0.1)	75.7(0.1)	70.4(0.1)	53.0(0.1)
	maximin	77.9(0.1)	84.2(0.1)	80.0(0.1)	57.6(0.1)
Setting 2	pooled	72.3(0.1)	75.5(0.1)	70.6(0.1)	53.2(0.1)
	maximin	77.9(0.2)	84.0(0.1)	80.1(0.1)	57.9(0.1)
Setting 3	pooled	72.5(0.2)	75.9(0.2)	70.9(0.2)	53.3(0.1)
	maximin	77.9(0.1)	84.2(0.1)	80.0(0.1)	57.6(0.1)
Setting 4	pooled	72.6(0.2)	75.9(0.3)	71.2(0.2)	53.7(0.1)
	maximin	77.7(0.2)	83.8(0.2)	80.2(0.2)	58.0(0.2)

Table 26: The PCD results (%, with standard errors in parenthesis) under the estimated maximin OTR and the pooled OTR for Scenario 2 when covariates are generated as in (ii).

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	81.0(<0.1)	97.6(<0.1)	92.0(0.1)	85.8(0.1)
	maximin	88.3(0.1)	92.1(0.1)	92.5(0.1)	87.7(0.1)
Setting 2	pooled	81.0(0.1)	97.2(0.1)	92.1(0.1)	85.9(0.1)
	maximin	88.4(0.2)	92.5(0.2)	92.2(0.2)	87.5(0.2)
Setting 3	pooled	80.9(0.1)	97.0(0.1)	92.1(0.1)	85.8(0.1)
	maximin	88.0(0.2)	92.9(0.1)	92.3(0.2)	87.9(0.2)
Setting 4	pooled	81.0(0.2)	96.2(0.1)	92.0(0.2)	85.7(0.2)
	maximin	87.8(0.3)	93.2(0.2)	91.2(0.3)	87.3(0.3)

Table 27: The PCD results (%, with standard errors in parenthesis) under the estimated maximin OTR and the pooled OTR for Scenario 1 when covariates are generated as in (iii).

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	66.2(0.1)	80.0(0.1)	81.2(0.1)	65.9(0.1)
	maximin	70.7(0.1)	84.7(<0.1)	87.2(<0.1)	69.9(0.1)
	pooled	66.3(0.1)	79.9(0.1)	81.1(0.1)	65.9(0.1)
	maximin	70.7(0.1)	84.7(0.1)	86.9(0.1)	69.5(0.1)
	pooled	66.6(0.1)	80.1(0.1)	80.9(0.1)	65.6(0.1)
	maximin	70.6(0.1)	84.7(0.1)	87.2(0.1)	69.8(0.1)
	pooled	66.7(0.2)	80.1(0.2)	80.7(0.2)	65.4(0.2)
	maximin	70.5(0.2)	84.6(0.1)	86.7(0.1)	69.1(0.2)

Table 28: The PCD results (%, with standard errors in parenthesis) under the estimated maximin OTR and the pooled OTR for Scenario 2 when covariates are generated as in (iii).

Testing group		First group	Second group	Third group	Fourth group
Setting 1	pooled	86.6(<0.1)	98.3(<0.1)	95.1(<0.1)	91.1(<0.1)
	maximin	91.5(0.1)	95.3(0.1)	95.5(0.1)	92.4(0.1)
Setting 2	pooled	86.6(0.1)	97.8(<0.1)	94.9(0.1)	91.1(0.1)
	maximin	91.1(0.2)	95.2(0.1)	95.0(0.1)	92.2(0.1)
Setting 3	pooled	86.8(0.1)	97.8(<0.1)	95.0(0.1)	91.0(0.1)
	maximin	91.4(0.2)	95.5(0.1)	95.0(0.1)	92.2(0.1)
Setting 4	pooled	86.9(0.1)	97.0(0.1)	94.6(0.1)	90.9(0.1)
	maximin	90.4(0.2)	95.1(0.1)	93.8(0.2)	91.3(0.2)

D. The schizophrenia data

Tarrier et al. (2004) conducted a multi-center, randomized controlled trial with 18-month follow-up , to examine the effects of cognitive-behavioral therapy (CBT) and supportive counselling (SC) on the outcomes of an early episode of schizophrenia. Patients were randomized to three treatment options, including the cognitive-behavioural therapy plus treatment as usual (CBT), supportive counselling plus treatment as usual (SC) and treatment as usual (TAU). The primary outcome, the Positive and Negative Syndromes Schedule (PANSS, Kay et al., 1987), was measured at baseline and the end of follow-up. Patients' durations of untreated psychosis, years of education and social functioning scores were also recorded at baseline.

As previous studies showed that both psychological treatment groups (CBT and SC) had a superior treatment effect compared to the control group (TAU), we focus on comparing two treatment arms: CBT ($A = 1$) and SC ($A = 0$) to determine individual OTRs. The reduction of PANSS score at the 18th month's visit is set as a patient's response Y. We consider two covariates: PANSS score at baseline ($X^{(1)}$) and log duration of untreated psychosis ($X^{(2)}$). Over 400 patients were initially enrolled in 3 treatment centres. Among them, only 165 finished the follow-up study and had completed records of the final response and baseline information. 85 of them received CBT or SC. As in Tarrier (2004), we classify 85 patients into 3 groups according to their treatment centres (Manchester, Liverpool and North Nottinghamshire). We first standardize the two covariates such that their sampling covariance matrix equals the identity matrix within each group and then jointly estimate c_0, β_{g1}, β_{g2} by the A-learning estimating equations as discussed previously. Estimators for β_{g1} and β_{g2} are given in Table 29.

Table 29: Estimators of groupwise OTR (standard errors in paranthesis) for the CBT study.

	Group 1	Group 2	Group 3
$\hat{\beta}_{g1}$	1.35(10.21)	1.17(11.52)	-20.71(13.27)
$\hat{\beta}_{g2}$	7.87(10.39)	-10.56(8.84)	3.45(9.14)

Differences of β_{gi} between different groups are not statistically significant. The large standard errors are due to the small sample size of each group. However, some of the estimated coefficients $\hat{\beta}_{gi}$'s among different groups are not even sign consistent, indicating potential existence of heterogeneity in optimal treatment regimes across different groups.

We adopt the leave-one-group-out cross validation procedure as done in the previous example. We report the estimated maximin OTR $\hat{d}_M$, the estimated pooled OTR $\hat{d}_P$, the OTR obtained based on random effects models $\hat{d}_R$, as well as the corresponding estimated value functions in Table 30. All three OTRs have similar value functions for Groups 1 and 2. However, for Group 3, value function under the maximin OTR is much higher than those under other OTRs.

Table 30: $\hat{d}_M$, $\hat{d}_P$, $\hat{d}_R$ and their estimated value functions.

Testing group	Group 1			Group 2			Group 3		
	$\hat{d}_M$	$\hat{d}_P$	$\hat{d}_R$	$\hat{d}_M$	$\hat{d}_P$	$\hat{d}_R$	$\hat{d}_M$	$\hat{d}_P$	$\hat{d}_R$
$\hat{c}$	0.33	-1.13	-1.25	4.25	4.52	3.26	2.72	1.83	1.05
$\hat{\beta}_1$	0.11	-0.45	-2.79	1.00	-2.42	-2.89	1.00	0.04	0.11
$\hat{\beta}_2$	-0.99	-0.45	-3.15	-0.07	-3.23	-5.06	-0.01	3.86	4.62
$\hat{\mathrm{E}}Y_g^{\star}(d)$	26.25	25.66	25.32	29.92	30.81	32.04	24.01	16.29	14.36

E. Proofs

E.1. Proof of theorem 1

We use the same notations in the proof of theorem 2. To prove theorem 1, we will show that for any fixed c, function $\mathrm{PCD}_g(\beta, c)$ can be presented as $\mathrm{E}\psi(\beta^T\beta_g, T)$ for some random variable T and function $\psi(\cdot,\cdot)$. In addition, $\psi(\cdot, t)$ is monotone increasing for fixed t. Then the assertion of theorem 1 follows by an application of lemma 4.

Without loss of generality, assume $\|\beta_g\|_2 = 1$ for all g. Recall that

$$\begin{aligned}\mathrm{PCD}_g(\beta, c) &= 1 - \mathrm{E}|I(X_g^T\beta > -c) - I(X_g^T\beta_g > -c_0)| = 1 - \mathrm{E}|I(X_g^T\beta > -c) - I(X_g^T\beta_g > -c_0)|^2 \\ &= 1 - \Pr(X_g^T\beta > -c) - \Pr(X_g^T\beta_g > -c_0) + 2\Pr(X_g^T\beta > -c, X_g^T\beta_g > -c_0). \qquad (4)\end{aligned}$$

Similar to the proof of theorem 2, we can show $\Pr(X_g^T\beta > -c)$ is constant as a function of $\{\beta \in \mathbb{R}^s :$

$\|\beta\|_2 = 1\}$. Similarly, under the condition that $\|\beta_1\|_2 = \cdots = \|\beta_G\|_2$, we can show $\Pr(X_g^T\beta_g > -c_0)$'s are the same for all g. Combining these with (4), we obtain

$$\mathrm{PCD}_g(\beta, c) = 2\Pr(X^T\beta > -c, X^T\beta_g > -c_0) + \xi(c), \tag{5}$$

for some function ξ independent of β. Recall $\beta_{(1)}^M = \arg\min_\beta \mathrm{PCD}_g(\beta, c)$. By (5), we have

$$\beta_{(1)}^M = \arg\max \min_g \Pr(X^T\beta > -c, X^T\beta_g > -c_0).$$

Therefore, it suffices to show for all β subject to the constraint $\|\beta\|_2 = 1$,

$$\min_g \Pr(X^T\beta > -c, X^T\beta_g > -c_0) \le \min_g \Pr(X^T\beta^M > -c, X^T\beta_g > -c_0). \tag{6}$$

It follows from lemma F.1 that for all g, there exists an orthogonal matrix Γ_g such that

$$\Gamma_g\beta = (1, 0, \ldots, 0)^T, \quad \Gamma_g\beta_g = (\beta^T\beta_g, \sqrt{1-(\beta^T\beta_g)^2}, 0, \ldots, 0)^T.$$

This implies

$$\begin{aligned} & \Pr(X^T\beta > -c, X^T\beta_g > -c_0) = \Pr(X^T\Gamma_g^T\Gamma_g\beta > -c, X^T\Gamma_g^T\Gamma_g\beta_g > -c_0) \\ = \; & \Pr(X^T\Gamma_g\beta > -c, X^T\Gamma_g\beta_g > -c_0) = \Pr\left(X^{(1)} > -c, X^{(1)}\beta_g^T\beta + X^{(2)}\sqrt{1-(\beta_g^T\beta)^2} > -c_0\right), \end{aligned} \tag{7}$$

where the second equality is due to the fact that X is spherically distributed (see Definition F.1), and that $X^{(1)}$ and $X^{(2)}$ are the first two components of the random vector X. It follows from theorem 2.6 in Fang et al. (1990) that

$$(X^{(1)}, X^{(2)}) \overset{d}{=} rd(U_1, U_2), \tag{8}$$

with $r = \|X\|_2$, $d \sim B(1, p/2-1)$, U_1 and U_2 uniformly distributed on the surface $u_1^2 + u_2^2 = 1$, where $B(p, q)$ stands for the Beta distribution with parameters p, q. The random variables r, d are independent of U_1 and U_2. Set $T = rd$. Combining this with (7) gives

$$\Pr(X^T\beta > -c, X^T\beta_g > -c_0) = \Pr\left\{TU_1 > -c, \left(\beta^T\beta_g U_1 + \sqrt{1-(\beta^T\beta_g)^2}U_2\right)T > -c_0\right\}. \tag{9}$$

For fixed $T = t$, the right-hand side of (9) is a function of $\beta^T\beta_g$ only. Moreover, it can be further represented as

$$\mathrm{E}[h(\beta^T\beta_g, t)|T = t] \equiv \mathrm{E}\left[\Pr\left\{tU_1 > -c, \left(\beta^T\beta_g U_1 + \sqrt{1-(\beta^T\beta_g)^2}U_2\right)t > -c_0\right\}|T = t\right],$$

by the independence between T and U_1, U_2. By lemma 4, it suffices to show that $h(\cdot, t)$ is monotonically increasing as a function of $\beta^T\beta_g$ for all t. When $t = 0$, this becomes trivial. Assume $t > 0$ and consider, separately, the cases where $\{c \le 0, c_0 \le 0\}$, $\{c > 0, c_0 \le 0\}$, $\{c \le 0, c_0 > 0\}$ and

$\{c > 0, c_0 > 0\}$. We only show h is monotonically increasing as a function of $\beta^T\beta_g$ when $c \leq 0, c_0 \leq 0$ and $c > 0, c_0 \leq 0$. In other two cases, the assertion can be established using similar arguments.

Case 1: $c \leq 0$, $c_0 \leq 0$. When either of c or c_0 is smaller than or equal to $-t$, $h = 0$. It suffices to consider when $-c/t = \sin\psi_1$, $-c_0/t = \sin\psi_2$ for $\psi_1, \psi_2 \in [0, \pi/2)$. Write $\beta^T\beta_g = \cos(\psi_3)$ for $\psi_3 \in [0, \pi]$. We now argue h is decreasing as ψ_3 increases. Recall

$$h = \Pr(U_1 > \sin\psi_1, \cos(\psi_3)U_1 + \sin(\psi_3)U_2 > \sin\psi_2). \tag{10}$$

Note that U_1 and U_2 can be presented as $U_1 = \sin\Theta$ and $U_2 = \cos\Theta$ for some random variable Θ uniformly distributed on $[0, 2\pi]$. With some calculation, RHS of (10) is equal to

$$\begin{aligned}
& \Pr(\sin\Theta > \sin\psi_1, \sin(\Theta + \psi_3) > \sin\psi_2) \qquad (11)\\
= \; & \Pr(\psi_1 < \Theta < \pi - \psi_1, \psi_2 - \psi_3 < \Theta < \pi - \psi_2 - \psi_3)\\
= \; & \frac{1}{2\pi}[\min(\pi - \psi_1, \pi - \psi_2 - \psi_3) - \max(\psi_1, \psi_2 - \psi_3)]_+\\
= \; & \begin{cases} \frac{1}{2\pi}[\pi - \psi_2 - \psi_3 - \max(\psi_1, \psi_2 - \psi_3)]_+, & \psi_2 > \psi_1,\\ \frac{1}{2\pi}[\min(\pi - \psi_1, \pi - \psi_2 - \psi_3) - \psi_1]_+, & \psi_2 \leq \psi_1, \end{cases}
\end{aligned}$$

where $[a]_+ \overset{\Delta}{=} \max(a, 0)$ for any real number a. Combining (10) with (11), we can see function h decreases as ψ_3 increases.

Case 2: $c > 0$, $c_0 \leq 0$. When $c_0 \leq -t$, $h = 0$. When $c \geq t$,

$$h = \Pr\left(U_1\beta^T\beta_g + U_2\sqrt{1 - (\beta^T\beta_g)^2} \geq -c_0/t\right) = \Pr(U_1 > -c_0/t),$$

which is a constant and does not change with $\beta^T\beta_g$. Hence, it suffices to consider cases where $c < t$, $c_0 > -t$. Assume $c/t = \sin\psi_1$, $-c_0/t = \sin\psi_2$ for some $\psi_1, \psi_2 \in [0, \pi/2)$, $\beta^T\beta_g = \cos\psi_3$ for $\psi_3 = [0, \pi]$. With some calculations, we can show that

$$\begin{aligned}
h \;=\; & \Pr(U_1 > -\sin\psi_1, \cos(\psi_3)U_1 + \sin(\psi_3)U_2 > \sin\psi_2)\\
= \; & \frac{1}{2\pi}[\pi - \psi_2 - \max(\psi_2, \psi_3 - \psi_1)]_+ + \frac{1}{2\pi}[\psi_1 + \psi_3 - \psi_2 - \pi]_+\\
= \; & \begin{cases} \frac{1}{2\pi}[\pi - \psi_2 - \max(\psi_2, \psi_3 - \psi_1)]_+, & \psi_2 > \psi_1,\\ \frac{1}{2\pi}\min(\pi - 2\psi_2, \max(\pi - \psi_2 - \psi_3 + \psi_1, 2\psi_1 - 2\psi_2)), & \psi_2 \leq \psi_1. \end{cases}
\end{aligned}$$

Hence, h is increasing as a function of $\beta^T\beta_g$. This completes the proof.

E.2. Proof of theorem 3

We assume $\|\beta_g\|_2 = 1$. In the proof of theorem 1, we have shown that for any β satisfying $\|\beta\|_2 = 1$, $\text{PCD}_g(\beta, c)$ is a function of $f(\beta^T\beta_g, c, c_0)$, which increases as a function of $\beta^T\beta_g$ for fixed c and c_0.

Hence, we obtain

$$\min_g \mathrm{PCD}_g(\beta^M_{(0)}, c) = \min_g f(\beta_g^T \beta^M_{(0)}, c, c_0) = f(\min_g \beta_g^T \beta^M_{(0)}, c, c_0).$$

Since $\min_g \beta_g^T \beta^M_{(0)} > 0$, it suffices to show that for any fixed $0 < \rho \leq 1$ and fixed c_0, the maximum of $f(\rho, c, c_0)$ as a function of c is achieved at $c = c_0/\rho$. Similar to (4) and (7), we can show

$$f(\rho, c, c_0) = 2\mathrm{Pr}(X^{(1)} > -c, X^{(1)}\rho + X^{(2)}\sqrt{1-\rho^2} > -c_0) - \mathrm{Pr}(X^{(1)} > -c) + q(c_0),$$

for some function $q(\cdot)$. Hence, it suffices to show that the maximum of

$$2\mathrm{Pr}(X^{(1)} > -c, X^{(1)}\rho + X^{(2)}\sqrt{1-\rho^2} > -c_0) - \mathrm{Pr}(X^{(1)} > -c)$$

as a function of c is achieved at $c = c_0/\rho$.

Due to the decomposition in (8), we have

$$\begin{aligned} & 2\mathrm{Pr}(X^{(1)} > -c, X^{(1)}\rho + X^{(2)}\sqrt{1-\rho^2} > -c_0) - \mathrm{Pr}(X^{(1)} > -c) \\ = \; & \mathrm{E}\left\{2I(tU_1 > -c, tU_1\rho + tU_2\sqrt{1-\rho^2} > -c_0) - I(tU_1 > -c)\right\} \triangleq \mathrm{E}\left\{h(t, \rho, c, c_0) | T = t\right\}. \end{aligned}$$

It suffices to show that the maximum of h is achieved at $c = c_0/\rho$ for fixed $t \geq 0$, ρ and c_0.

Note that

$$h(t, \rho, c, c_0) = 2\mathrm{E}I(tU_1 > -c, tU_1\rho + tU_2\sqrt{1-\rho^2} > -c_0) - \mathrm{E}I(tU_1 > -c).$$

When $t = 0$,

$$h = 2I(c < 0, c_0 < 0) - I(c < 0).$$

Therefore, the maximum of h is achieved for any c such that $\mathrm{sgn}(c) = \mathrm{sgn}(c_0)$, where sgn stands for the sign function with $\mathrm{sgn}(c_0) = 1$ for all $c_0 \geq 0$ and $\mathrm{sgn}(c_0) = -1$ for all $c_0 < 0$. Since $0 < \rho \leq 1$, we have $\mathrm{sgn}(c_0/\rho) = \mathrm{sgn}(c_0)$. This verifies that c_0/ρ is the maximizer of h when $t = 0$.

When $t > 0$, note that $h(t, \rho, c, c_0) = g(\rho, c/t, c_0/t)$ where

$$g(\rho, c^*, c_0^*) = \mathrm{E}\left\{2I(U_1 > -c^*, U_1\rho + U_2\sqrt{1-\rho^2} > -c_0^*) - I(U_1 > -c^*)\right\}.$$

As a result, we only need to show that the maximum of g is achieved at $c^* = c_0^*/\rho$ for fixed c_0^* and ρ.

Assume $\rho = \cos(\psi_1)$ for some $\psi_1 \in (0, \pi/2)$. We focus on the case where $c_0^* = -\cos(\psi_2)$ for some $\psi_2 \in (0, \pi)$. When $|c_0^*| \geq 1$, the assertion that c^* is the maximizer can be easily proven. With some calculation, function g is equal to

$$g(\rho, c^*, c_0^*) = \begin{cases} 0, & c^* \leq -1, \\ \frac{[\min(\psi_1+\psi_2, \psi_3) - \max(\psi_1-\psi_2, -\psi_3)]_+ + [\psi_1+\psi_2+\psi_3-2\pi]_+ - \psi_3}{\pi}, & c^* = -\cos(\psi_3), \psi_3 \in (0, \pi), \\ 2\psi_2/\pi - 1, & c^* \geq 1. \end{cases}$$

Consider separately the cases where $\cos(\psi_2)/\cos(\psi_1) > 1$, $\cos(\psi_2)/\cos(\psi_1) < 1$, and $|\cos(\psi_2)/\cos(\psi_1)| \le 1$. We show that g is maximized at $c^* = c_M^* = -\cos(\psi_2)/\cos(\psi_1)$ in all these three cases. The proof is thus completed.

Case 1: $\cos(\psi_2)/\cos(\psi_1) > 1$. In this case, $c_M^* \le -1$. Since $\psi_1 \in (0, \pi/2)$, this means $\psi_2 \in (0, \pi/2)$ and hence

$$g(\rho, c^*, c_0^*) = \frac{2\psi_2}{\pi} - 1 < 0 = g(\rho, c_M^*, c_0^*), \qquad \forall c^* \ge 1.$$

Thus, it suffices to show that $g \le 0$ for $c^* = -\cos(\psi_3)$ for some $\psi_3 \in (0, \pi)$. When $\cos(\psi_2) > \cos(\psi_1)$, we have $\psi_1 > \psi_2$ and hence

$$\max(\psi_1 - \psi_2, -\psi_3) = \psi_1 - \psi_2 > 0.$$

Besides, since $\psi_1, \psi_2 \in (0, \pi/2)$, $\psi_3 \in (0, \pi)$, we have $\psi_1 + \psi_2 + \psi_3 \le 2\pi$ and hence

$$[\psi_1 + \psi_2 + \psi_3 - 2\pi]_+ = 0.$$

Therefore, when $c^* = -\cos(\psi_3)$,

$$\begin{aligned} g &= \frac{1}{\pi}\left([\min(\psi_1 + \psi_2, \psi_3) - \max(\psi_1 - \psi_2, -\psi_3)]_+ + [\psi_1 + \psi_2 + \psi_3 - 2\pi]_+ - \psi_3\right) \\ &\le \frac{1}{\pi}\left([\min(\psi_1 + \psi_2, \psi_3)]_+ - \psi_3\right) = \frac{1}{\pi}(\psi_3 - \psi_3) = 0. \end{aligned}$$

This shows that g is maximized at $c^* = c_M^*$ when $\cos(\psi_2)/\cos(\psi_1) > 1$.

Case 2: $\cos(\psi_2)/\cos(\psi_1) < -1$. In this case, we have $\cos(\psi_2) + \cos(\psi_1) < 0$, and hence $c_M^* = -\cos(\psi_2)/\cos(\psi_1) > 1$. Besides,

$$\cos(\pi - \psi_2) = -\cos(\psi_2) > \cos(\psi_1). \tag{12}$$

Equation (12) implies $\psi_1 + \psi_2 > \pi$ and hence $2\psi_2/\pi - 1 > 0$. Thus,

$$g(\rho, c_M^*, c_0^*) > 0 = g(\rho, c^*, c_0^*), \qquad \forall c^* \le -1. \tag{13}$$

It suffices to show $g(\rho, c^*, c_0^*) \le 2\psi_2/\pi - 1$ for $c^* = -\cos(\psi_3)$ for some ψ_3. Since $\psi_1 + \psi_2 > \pi \ge \psi_3$, $\psi_1 \in (0, \pi/2)$, we obtain $\psi_2 \in (\pi/2, \pi)$ and $\psi_2 - \psi_1 > 0$. Therefore,

$$\begin{aligned} & g(\rho, -\cos(\psi_3), -\cos(\psi_2)) \le \frac{1}{\pi}\left([\psi_3 - \psi_1 + \psi_2]_+ + [\psi_1 + \psi_2 + \psi_3 - 2\pi]_+ - \psi_3\right) \\ =\; & \frac{1}{\pi}(\psi_3 - \psi_1 + \psi_2 - \psi_3 + [\psi_1 + \psi_2 + \psi_3 - 2\pi]_+) = \frac{1}{\pi}(\psi_2 - \psi_1 + [\psi_1 + \psi_2 + \psi_3 - 2\pi]_+) \\ \le\; & \frac{1}{\pi}(\psi_2 - \psi_1 + [\psi_1 + \psi_2 - \pi]_+) = \frac{1}{\pi}(\psi_2 - \psi_1 + \psi_1 + \psi_2 - \pi) = \frac{2}{\pi}\psi_2 - 1. \end{aligned}$$

This together with (13) suggest that g is maximized at $c^* = c_M^*$, when $\cos(\psi_2)/\cos(\psi_1) < -1$.

Case 3: Finally we show that g is maximized at $c^* = c^*_M$ when $|\cos(\psi_2)/\cos(\psi_1)| \leq 1$. Since $\psi_1 \in (0, \pi/2)$, this suggests $\psi_1 \leq \psi_2 \leq \pi - \psi_1$. Hence, we have $\psi_1 + \psi_2 + \psi_3 \leq \pi + \psi_3 \leq 2\pi$ and

$$[\psi_1 + \psi_2 + \psi_3 - 2\pi]_+ = 0. \tag{14}$$

We assume $\cos(\psi_2)/\cos(\psi_1) = \cos(\psi_0)$ for some $\psi_0 \in [0, \pi]$. We aim to show that g is maximized at $c^* = -\cos(\psi_3)$ for $\psi_3 = \psi_0$.

For any $\pi > \psi_3 \geq \psi_0$, we have $\cos(\psi_3) \leq \cos(\psi_0)$ and hence $\cos(\psi_3)\cos(\psi_1) \leq \cos(\psi_2)$. This implies

$$\cos(\psi_1 + \psi_3) = \cos(\psi_1)\cos(\psi_3) - \sin(\psi_1)\sin(\psi_3) \leq \cos(\psi_2),$$

which further yields

$$\psi_1 + \psi_3 \geq \psi_2, \qquad \forall \psi_3 \geq \psi_0. \tag{15}$$

By (14) and (15), we have for any $\psi_3 \geq \psi_0$,

$$\begin{aligned} g(\rho, -\cos(\psi_3), -\cos(\psi_2)) &= \frac{1}{\pi}\left([\min(\psi_1 + \psi_2, \psi_3) - \max(\psi_1 - \psi_2, -\psi_3)]_+ - \psi_3\right) \qquad (16)\\ = \frac{1}{\pi}\left([\min(\psi_1 + \psi_2, \psi_3) - \psi_1 + \psi_2]_+ - \psi_3\right) &\leq \frac{1}{\pi}\left([\psi_3 + \psi_2 - \psi_1]_+ - \psi_3\right) = \frac{1}{\pi}(\psi_2 - \psi_1), \end{aligned}$$

where the last equality is due to that $\psi_2 \geq \psi_1$ and hence $\psi_3 + \psi_2 - \psi_1 \geq 0$.

For any $0 < \psi_3 \leq \psi_0$, we have $\cos(\psi_3) \geq \cos(\psi_0)$ and hence $\cos(\psi_3)\cos(\psi_1) \geq \cos(\psi_2)$. Therefore, we obtain

$$\cos(\psi_3 - \psi_1) = \cos(\psi_1)\cos(\psi_3) + \sin(\psi_1)\sin(\psi_3) \geq \cos(\psi_2).$$

This suggests

$$\psi_1 + \psi_2 \geq \psi_3, \qquad \forall \psi_3 \leq \psi_0. \tag{17}$$

By (14) and (17), we have for any $\psi_3 \leq \psi_0$,

$$\begin{aligned} g(\rho, -\cos(\psi_3), -\cos(\psi_2)) &\leq \frac{1}{\pi}\left([\psi_3 - \max(\psi_1 - \psi_2, -\psi_3)]_+ - \psi_3\right) \qquad (18)\\ \leq \frac{1}{\pi}[\psi_3 - \psi_1 + \psi_2]_+ - \frac{1}{\pi}\psi_3 &= \frac{1}{\pi}(\psi_2 - \psi_1). \end{aligned}$$

Set $\psi_3 = \psi_0$. It follows from (15), (17) and $\psi_2 \geq \psi_1$ that

$$\begin{aligned} g(\rho, -\cos(\psi_0), -\cos(\psi_2)) &= \frac{1}{\pi}\left([\min(\psi_1 + \psi_2, \psi_0) - \max(\psi_1 - \psi_2, -\psi_0)]_+ - \psi_0\right) \qquad (19)\\ = \frac{1}{\pi}\left([\psi_0 - \psi_1 + \psi_2]_+ - \psi_0\right) &= \frac{1}{\pi}\left([\psi_0 + \psi_2 - \psi_1]_+ - \psi_0\right) = \frac{1}{\pi}(\psi_2 - \psi_1). \end{aligned}$$

Combining (19) together with (16) and (18), we have shown that

$$g(\rho, -\cos(\psi_0), -\cos(\psi_2)) \geq g(\rho, c^*, -\cos(\psi_2)), \tag{20}$$

for all $|c^*| \leq 1$.

By (19), since $\psi_2 \geq \psi_1$ and $g = 0$ when $c^* \leq -1$, (20) also holds for all $c^* \leq -1$.

It remains to show (20) holds for all $c^* \geq 1$. Since $\psi_2 \leq \pi - \psi_1$, when $c^* \geq 1$, we have

$$g = \frac{2\psi_2}{\pi} - 1 \leq \frac{\psi_2 + \pi - \psi_1}{\pi} - 1 \leq \frac{\psi_2 - \psi_1}{\pi}.$$

It follows from (19) that $g(\rho, -\cos(\psi_0), -\cos(\psi_2)) \geq g(\rho, c^*, -\cos(\psi_2))$ when $c^* \geq 1$. The proof is then completed.

E.3. Proof of theorem 4

For any g, define $\rho_g = \beta_g^T \beta$. It follows from lemma F.1 and the decomposition in (8) that

$$\begin{aligned} \mathrm{VD}_g(\beta, c) = \mathrm{E}(X^T\beta_g + c_0) I(X^T\beta + c > 0) = \mathrm{E}\left\{ \|\beta_g\|_2 \left(X^{(1)}\rho_g + X^{(2)}\sqrt{1-\rho_g^2} \right) I(X^{(1)} > -c) \right\} \\ + \quad c_0 \mathrm{Pr}(X^{(1)} > -c) = \mathrm{E}\{\|\beta_g\|_2 X^{(1)} \rho_g I(X^{(1)} > -c)\} + c_0 \mathrm{Pr}(X^{(1)} > -c) = \mathrm{E}(\beta^T\beta_g X^{(1)} + c_0) I(X^{(1)} > -c). \end{aligned}$$

Using similar arguments in the proof of theorem 3, it suffices to show for any $t > 0$, and fixed c_0, the function

$$\mathrm{E}\{(tX^{(1)} + c_0) I(X^{(1)} > -c)\}$$

is maximized at $c = c_0/t$. However, this is immediate to see since $tx + c_0 > 0$ when $x > -c_0/t$, and $tx + c_0 \leq 0$ when $x \leq c_0/t$. The proof is thus completed.

E.4. Proof of Lemma 3

Define $F(\beta) = \min_g \beta^T\beta_g$. We present the following lemmas before proving lemma 3.

Lemma E.1. *For any vector β_0, consider the set*

$$K(\beta_0) = \{g = 1, \dots, G : F(\beta_0) = \beta_0^T \beta_g\}.$$

Then there exists an $\varepsilon > 0$ such that for all $\|\beta - \beta_0\|_2 \leq \varepsilon$, we have

$$F(\beta) = \min_{g \in K(\beta_0)} \beta^T \beta_g.$$

Lemma E.2. *If β^M is the solution of $\max_{\|\beta\|_2 = 1} F(\beta)$, then there exists a unique nonempty subset $K_0 \subseteq [1, \dots, G]$ such that*

$$\beta^M \in E_{K_0}(B).$$

Besides, we have

$$\min_{j\in K_0^c} \beta_j^T\beta^M > \beta_k^T\beta^M, \qquad \forall k \in K_0.$$

LEMMA E.3. *Define $F_{K_0}(\beta) = \min_{g\in K_0} \beta^T\beta_g$ for any $K_0 \subseteq [1,\dots,G]$. Then for any unit vector $\beta \in C(B_{K_0})$, and any unit vector β_0 such that $\|\beta_0 - \beta\|_2 < \varepsilon$, there exists another unit length vector $\beta' \in C(B_{K_0})$ such that $\|\beta' - \beta\|_2 < \varepsilon$, and $F_{K_0}(\beta') \geq F_{K_0}(\beta_0)$.*

Proof of lemma 3: Since $G_0 > 0$, we have $\beta_{(0)}^M = \beta^M$ where $\beta^M = \arg\max_{\|\beta\|_2\leq 1} F(\beta)$. We first show there exists a subset $K_0 \subseteq [1,\dots,G]$ such that

$$\beta^M = \mathrm{E}^\star_{K_0}(B), \tag{21}$$

with

$$\min_{j\in K_0^c} \beta_j^T\beta^M > \beta_k^T\beta^M, \qquad \forall k \in K_0. \tag{22}$$

Lemma E.2 asserts that there exists a unique set $K_0 \subseteq [1,\dots,G]$ such that $\beta^M \in \mathrm{E}_{K_0}(B)$ and (22) holds. For this K_0, it follows from lemma 2 that $\mathrm{E}^\star_{K_0}(B)$ exists.

Define

$$\beta^+ = \frac{\beta^M + \delta\mathrm{E}^\star_{K_0}(B)}{\{1+\delta^2+2\delta{\beta^M}^T\mathrm{E}^\star_{K_0}(B)\}^{1/2}}.$$

Note that $\|\beta^+\|_2 = 1$.

By lemma E.1, for sufficiently small $\delta > 0$, we have

$$F(\beta^+) = \min_{g\in K_0} \frac{\beta_g^T\beta^M + \delta\beta_g^T\mathrm{E}^\star_{K_0}(B)}{\{1+\delta^2+2\delta{\beta^M}^T\mathrm{E}^\star_{K_0}(B)\}^{1/2}}.$$

It follows by the definition of $\mathrm{E}^\star_{K_0}(B)$ and $\beta^M \in \mathrm{E}_{K_0}(B)$ that $\beta_g^T\beta^M \leq \beta_g^T\mathrm{E}^\star_{K_0}(B)$, for all $g \in K_0$. Therefore, we have

$$F(\beta^+) \geq \min_{g\in K_0} \frac{(1+\delta)\beta_g^T\beta^M}{\{1+\delta^2+2\delta{\beta^M}^T\mathrm{E}^\star_{K_0}(B)\}^{1/2}}. \tag{23}$$

Assume $\beta^M \neq \mathrm{E}^\star_{K_0}(B)$, we have $(\beta^M)^T\mathrm{E}^\star_{K_0}(B) < 1$ and hence RHS of (23) is strictly larger than

$$\min_{g\in K_0} \frac{(1+\delta)\beta_g^T\beta^M}{(1+\delta^2+2\delta)^{1/2}} = \min_{g\in K_0} \beta_g^T\beta^M = F(\beta^M). \tag{24}$$

Combining (24) together with (23), we obtain $F(\beta^+) > F(\beta^M)$. However, this contradicts the definition of β^M. Assertion (21) hence follows.

We next show if there exists some non-empty set $K_0 \subseteq [1,\dots,G]$, $\beta_0 = \mathrm{E}^\star_{K_0}(B)$ such that

$$\min_{j\in K_0^c} \beta_j^T\beta_0 > \beta_k^T\beta_0, \qquad \forall k \in K_0, \tag{25}$$

and the column vectors in B_{K_0} are linearly independent, then a sufficient and necessary condition to establish $\beta_{(0)}^M = \beta_0$ is that

$$e_k^T (B_{K_0}^T B_{K_0})^{-1} e \geq 0, \qquad \forall k = 1, \dots, |K_0|, \tag{26}$$

where e_k is a basis vector with the kth component equal to 1, and other elements 0.

Note that $\beta_{(0)}^M = \arg\max_{\|\beta\|_2 \leq 1} \min_g \beta^T \beta_g$. The above optimization problem is concave. As a result, in order to show $\beta_{(0)}^M = \beta_0$, it suffices to show that for any β within the ε-neighborhood of β_0, we have $F(\beta) \leq F(\beta_0)$. By (25) and lemma E.1, for sufficiently small ε and β such that $\|\beta - \beta_0\|_2 \leq \varepsilon$, we have

$$F(\beta) = \min_{g \in K_0} \beta^T \beta_g.$$

To show β_0 is the maximizer of $F(\beta)$, by lemma E.3, we only need to show

$$F_{K_0}(\beta) \leq F_{K_0}(\beta_0), \qquad \forall \beta \in C(B_{K_0}), \quad \|\beta - \beta_0\|_2 \leq \varepsilon, \quad \|\beta\|_2 = 1.$$

Since vectors in B_{K_0} are linearly independent, for any $\beta \in C(B_{K_0})$ such that $\|\beta\|_2 = 1$, there exists a unique ω such that

$$\beta = B_{K_0} \omega,$$

with $\omega^T B_{K_0}^T B_{K_0} \omega = 1$. Similarly we present $\beta_0 = B_{K_0} \omega_0$ with $\omega_0^T B_{K_0}^T B_{K_0} \omega_0 = 1$. Since the column vectors in B_{K_0} are linearly independent, we have $e \in C(B_{K_0}^T)$. By definition, we have $\beta_0 = \mathrm{E}_{K_0}^{\star}(B)$. It follows from lemma 2 that

$$\beta_0 = \{e^T (B_{K_0}^T B_{K_0})^{-1} e\}^{-1/2} B_{K_0} (B_{K_0}^T B_{K_0})^{-1} e,$$

and hence

$$\omega_0 = \{e^T (B_{K_0}^T B_{K_0})^{-1} e\}^{-1/2} (B_{K_0}^T B_{K_0})^{-1} e. \tag{27}$$

For any ω such that $\omega^T B_{K_0}^T B_{K_0} \omega = 1$, it follows from Cauchy-Swartz inequality that

$$\omega^T B_{K_0}^T B_{K_0} \omega_0 \leq \omega_0^T B_{K_0}^T B_{K_0} \omega_0,$$

or equivalently,

$$(\omega - \omega_0)^T B_{K_0}^T B_{K_0} \omega_0 \leq 0. \tag{28}$$

We first show the sufficiency of (26). Assume for now (26) holds. It follows from (27) that all elements in ω_0 are nonnegative. By (28), this means for any $\beta = B_{K_0} \omega$ with $\|\beta\|_2 = 1$, at least one element in the vector $B_{K_0}^T B_{K_0} (\omega - \omega_0)$ must be smaller than or equal to 0. Note that

$$B_{K_0}^T (\beta - \beta_0) = B_{K_0}^T B_{K_0} (\omega - \omega_0). \tag{29}$$

It follows from (28) and (29) that for any $\beta \in C(B_{K_0})$ with unit L_2 norm, there exists some $k \in [1, \ldots, |K_0|]$ such that

$$e_k^T B_{K_0}^T (\beta - \beta_0) \leq 0. \tag{30}$$

Since β_0 is the optimal equicorrelated point, $e_k^T B_{K_0}^T \beta_0$ remains the same for all $k = 1, \ldots, |K_0|$. This together with (30) suggests that for any $\beta \in C(B_{K_0})$ with unit L_2 norm,

$$\min_k e_k^T B_{K_0}^T \beta \leq \min_k e_k^T B_{K_0}^T \beta_0,$$

or equivalently, $F(\beta) \leq F(\beta_0)$. The sufficiency thus follows.

To show the necessity, note that when at least one element in ω_0 is negative, we can construct some vector b with all positive elements such that $\omega_0^T b < 0$. Define

$$\delta = -\frac{2\omega_0^T b}{b^T (B_{K_0}^T B_{K_0})^{-1} b} (B_{K_0}^T B_{K_0})^{-1} b.$$

With some calculation, we have

$$\begin{aligned} & \|B_{K_0}(\delta + \omega_0)\|_2^2 = \delta^T B_{K_0}^T B_{K_0} \delta + 2\delta^T B_{K_0}^T B_{K_0} \omega_0 + \omega_0^T B_{K_0}^T B_{K_0} \omega_0 \\ = \quad & \frac{4(\omega_0^T b)^2}{b^T (B_{K_0}^T B_{K_0})^{-1} b} - \frac{4(\omega_0^T b)^2}{b^T (B_{K_0}^T B_{K_0})^{-1} b} + 1 = 1. \end{aligned}$$

This implies the vector $\beta = B_{K_0}^T (\delta + \omega_0)$ satisfies the L_2 unit norm constraint. Besides,

$$B_{K_0}(\beta - \beta_0) = B_{K_0}^T B_{K_0} \delta = -\frac{2\omega_0^T b}{b^T (B_{K_0}^T B_{K_0})^{-1} b} b. \tag{31}$$

Since all elements in b are positive and $\omega_0^T b < 0$, each element in RHS of (31) is positive. This implies

$$\min_k e_k^T B_{K_0}^T \beta > \min_k e_k^T B_{K_0}^T \beta_0,$$

and hence $F(\beta) > F(\beta_0)$. Therefore, we've reached a contradiction. The necessity thus follows.

E.5. Proof of theorem 5

Assume $F_0 < 0$, we have for any $\beta \neq 0$,

$$\begin{aligned} \min_g \beta^T \hat{\beta}_g &\leq \min_g [\beta^T \beta_g + \beta^T (\hat{\beta}_g - \beta_g)] \leq \min_g \beta^T \beta_g + \max_g \|\beta\|_2 \|\hat{\beta}_g - \beta_g\|_2 \\ &\leq \|\beta\|_2 \left(\min_g \frac{\beta^T \beta_g}{\|\beta\|_2} + \max_g \|\hat{\beta}_g - \beta_g\|_2 \right) = \|\beta\|_2 \left(F_0 + \max_g \|\hat{\beta}_g - \beta_g\|_2 \right). \end{aligned}$$

It follows from Condition (C1) that

$$\Pr\left(\max_g \|\hat{\beta}_g - \beta_g\|_2 > -\frac{F_0}{2} \right) \to 0.$$

Hence, we have

$$\Pr\left(\sup_{\beta\neq 0}\min_{g}\beta^T\hat{\beta}_g < 0\right) \to 1.$$

This implies with probability tending to 1, 0 is the estimated maximin coefficients. Therefore, $\hat{\beta}^M$ is consistent when $F_0 < 0$.

Consider the case when $F_0 > 0$. Define

$$\tilde{\beta}^M = [e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e]^{-1/2}\hat{B}_{K_0}(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e.$$

In order to show with probability tending to 1, $\hat{\beta}^M = \tilde{\beta}^M$, it follows from lemma 3 that we need to show with probability tending to 1, (i) the matrix $\hat{B}_{K_0}^T\hat{B}_{K_0}$ is invertible, (ii) $\min_{g\in K_0^c}\hat{\beta}_g^T\tilde{\beta}^M > \max_{g\in K_0}\hat{\beta}_g^T\tilde{\beta}^M$, and (iii) each element in the vector $(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e$ is nonnegative. We now break the proof into five steps. In the first step, we show $\tilde{\beta}^M$ is consistent with respect to $\beta_{(0)}^M$ and establish its convergence rate. In the next three steps, we verify (i)-(iii), respectively. Convergence rate of $\hat{\beta}^M$ is the same as that of $\tilde{\beta}^M$. Finally, we show the convergence rate of $\hat{c}^M$.

Step 1: Note that $\tilde{\beta}^M - \beta_{(0)}^M$ is equal to

$$[e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e]^{-1/2}\hat{B}_{K_0}(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e - [e^T(B_{K_0}^TB_{K_0})^{-1}e]^{-1/2}B_{K_0}(B_{K_0}^TB_{K_0})^{-1}e.$$

We decompose it as $I_1 + I_2 + I_3$ where

$$\begin{aligned}
I_1 &= \left([e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e]^{-1/2} - [e^T(B_{K_0}^TB_{K_0})^{-1}e]^{-1/2}\right)\hat{B}_{K_0}(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e,\\
I_2 &= [e^T(B_{K_0}^TB_{K_0})^{-1}e]^{-1/2}\left(\hat{B}_{K_0}(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e - B_{K_0}(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e\right),\\
I_3 &= [e^T(B_{K_0}^TB_{K_0})^{-1}e]^{-1/2}B_{K_0}\left((\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1} - (B_{K_0}^TB_{K_0})^{-1}\right)e.
\end{aligned}$$

Recall that $r_n^{(1)}$ the convergence rate of $\max_{g\in K_0}\|\hat{\beta}_g - \beta_g\|_2$. In the following, we argue each $\|I_j\|_2$ is of the order $= O_p(r_n^{(1)})$, for $j = 1, 2, 3$. We first prove $\|I_1\|_2 = O_p(r_n^{(1)})$. Before that, we show

$$\|\hat{B}_{K_0}^T\hat{B}_{K_0} - B_{K_0}^TB_{K_0}\|_2 = O_p(r_n^{(1)}) \xrightarrow{P} 0. \tag{32}$$

Note that the matrix $\hat{B}_{K_0}^T\hat{B}_{K_0} - B_{K_0}^TB_{K_0}$ is symmetric, it follows from lemma F.3 that the LHS in (32) is smaller than $\|\hat{B}_{K_0}^T\hat{B}_{K_0} - B_{K_0}^TB_{K_0}\|_\infty$. Besides, we have

$$\begin{aligned}
&\|\hat{B}_{K_0}^T\hat{B}_{K_0} - B_{K_0}^TB_{K_0}\|_\infty \le \max_{g\in K_0}\sum_{j\in K_0}|\hat{\beta}_g^T\hat{\beta}_j - \beta_g^T\beta_j| \qquad (33)\\
\le\ & \max_{g\in K_0}\sum_{j\in K_0}\left(|\hat{\beta}_g^T\hat{\beta}_j - \beta_g^T\hat{\beta}_j| + |\beta_g^T\hat{\beta}_j - \beta_g^T\beta_j|\right)\\
\le\ & \max_{g\in K_0}\left(\|\hat{\beta}_g\|_2 + \|\beta_g\|_2\right)\max_{g\in K_0}\sum_{j\in K_0}\left(\|\hat{\beta}_g - \beta_g\|_2 + \|\hat{\beta}_j - \beta_j\|_2\right) = O(r_n^{(1)}).
\end{aligned}$$

Therefore, (32) is proven. Note that

$$\begin{aligned}\lambda_{\min}(\hat{B}_{K_0}^T\hat{B}_{K_0}) &= \min_{\|a\|_2=1} a^T\hat{B}_{K_0}^T\hat{B}_{K_0}a \geq \min_{\|a\|_2=1} a^T B_{K_0}^T B_{K_0}a - \max_{\|a\|_2=1}|a^T(\hat{B}_{K_0}^T\hat{B}_{K_0} - B_{K_0}^T B_{K_0})a| \\ &\geq \lambda_{\min}(B_{K_0}^T B_{K_0}) - \|\hat{B}_{K_0}^T\hat{B}_{K_0} - B_{K_0}^T B_{K_0}\|_2. \qquad (34)\end{aligned}$$

Since the matrix B_{K_0} is of full column rank, the matrix $B_{K_0}^T B_{K_0}$ is invertible and hence $\lambda_{\min}(B_{K_0}^T B_{K_0}) > 0$. This together with (34) implies that with probability tending to 1,

$$\liminf \lambda_{\min}(\hat{B}_{K_0}^T\hat{B}_{K_0}) > 0 \text{ and } \lambda_{\max}\left((\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}\right) = O(1). \qquad (35)$$

Similarly we can show that with probability tending to 1,

$$\liminf \lambda_{\min}\left((\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}\right) > 0 \text{ and } \lambda_{\max}(\hat{B}_{K_0}^T\hat{B}_{K_0}) = O(1). \qquad (36)$$

It follows from Cauchy-Schwarz inequality that

$$\|\hat{B}_{K_0}(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e\|_2 \leq \sqrt{\|\hat{B}_{K_0}(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e\|_2^2} = \sqrt{\|e\|_2^2\lambda_{\max}\left((\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}\right)}. \qquad (37)$$

This together with (35) suggests $\|\hat{B}_{K_0}(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e\|_2 = O(1)$, with probability tending to 1.

Observe that $\|I_1\|_2$ is bounded from above by

$$|[e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e]^{-1/2} - [e^T(B_{K_0}^T B_{K_0})^{-1}e]^{-1/2}|\|\hat{B}_{K_0}(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e\|_2.$$

To show $\|I_1\|_2 = O_p(r_n^{(1)})$, it suffices to show

$$|[e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e]^{-1/2} - [e^T(B_{K_0}^T B_{K_0})^{-1}e]^{-1/2}| = O_p(r_n^{(1)}). \qquad (38)$$

LHS of (38) can be represented as

$$\frac{|\left(e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e\right)^{1/2} - \left(e^T(B_{K_0}^T B_{K_0})^{-1}e\right)^{1/2}|}{\left(e^T(B_{K_0}^T B_{K_0})^{-1}e\right)^{1/2}\left(e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e\right)^{1/2}}. \qquad (39)$$

Note that

$$e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e \geq \|e\|_2^2\lambda_{\min}\left((\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}\right).$$

This together with (36) implies that with probability tending to 1, the denominator in (39) is uniformly greater than some constant $c > 0$, for sufficiently large n.

Hence it suffices to show

$$\left(e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e\right)^{1/2} - \left(e^T(B_{K_0}^T B_{K_0})^{-1}e\right)^{1/2} = O_p(r_n^{(1)}). \qquad (40)$$

It follows from (36) that

$$\liminf\left\{\left(e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e\right)^{1/2} + \left(e^T(B_{K_0}^T B_{K_0})^{-1}e\right)^{1/2}\right\} > 0,$$

with probability tending to 1.

Note that

$$\begin{aligned} & \left(e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e - e^T(B_{K_0}^T B_{K_0})^{-1}e\right) \\ = & \left\{\left(e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e\right)^{1/2} - \left(e^T(B_{K_0}^T B_{K_0})^{-1}e\right)^{1/2}\right\}\left\{\left(e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e\right)^{1/2} + \left(e^T(B_{K_0}^T B_{K_0})^{-1}e\right)^{1/2}\right\}. \end{aligned}$$

It suffices to show

$$\left(e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e - e^T(B_{K_0}^T B_{K_0})^{-1}e\right) = O_p(r_n^{(1)}). \tag{41}$$

Note that the absolute value of the LHS of (41) can be bounded from above by

$$\|e\|_2^2\|(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1} - (B_{K_0}^T B_{K_0})^{-1}\|_2. \tag{42}$$

It follows from (32) and (35) that

$$\begin{aligned} & \|(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1} - (B_{K_0}^T B_{K_0})^{-1}\|_2 \qquad (43) \\ \le & \|(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}\|_2\|\hat{B}_{K_0}^T\hat{B}_{K_0} - B_{K_0}^T B_{K_0}\|_2\|(B_{K_0}^T B_{K_0})^{-1}\|_2 = O_p(r_n^{(1)}). \end{aligned}$$

Combining (42) together with (43), we obtain (41). Hence, we have shown $\|I_1\|_2 = O_p(r_n^{(1)})$.

Next we show $\|I_2\|_2 = O_p(r_n^{(1)})$. Note that $\|I_2\|_2$ can be bounded from above by

$$[e^T(B_{K_0}^T B_{K_0})^{-1}e]^{-1/2}\|\hat{B}_{K_0} - B_{K_0}\|_2\|(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e\|_2. \tag{44}$$

Similar to (37), we can show $\|(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e\|_2 = O_p(1)$. By (44), it suffices to show

$$\|\hat{B}_{K_0} - B_{K_0}\|_2 = O_p(r_n^{(1)}). \tag{45}$$

However, it is immediate to see (45) holds by definitions of $\hat{B}_{K_0}$ and B_{K_0}.

As for $\|I_3\|_2$, we can similarly show

$$\|I_3\|_2 \le [e^T(B_{K_0}^T B_{K_0})^{-1}e]^{-1/2}\|B_{K_0}\|_2\|(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1} - (B_{K_0}^T B_{K_0})^{-1}\|_2\|e\|_2. \tag{46}$$

By (32), the RHS of (46) is $O_p(r_n^{(1)})$. This implies $\|I_3\|_2 = O_p(r_n^{(1)})$. Under Condition (C1), we have $r_n^{(1)} \to 0$. Therefore, $\tilde{\beta}^M$ is consistent.

Step 2: Since B_{K_0} is of full column rank, the matrix $B_{K_0}^T B_{K_0}$ is invertible and hence $\lambda_{\min}(B_{K_0}^T B_{K_0}) > 0$. It follows from (32) and (34) that the matrix $\hat{B}_{K_0}^T\hat{B}_{K_0}$ is invertible with probability tending to 1.

Step 3: In Step 1, we have shown $\tilde{\beta}^M$ is consistent to $\beta_{(0)}^M$. Under Condition (C1), all estimators are consistent. Hence, we have

$$\begin{aligned} \min_{g\in K_0^c}\hat{\beta}_g^T\tilde{\beta}^M & \ge \min_{g\in K_0^c}\beta_g^T\beta_{(0)}^M - \max_{g\in K_0^c}\|\tilde{\beta}_g\|_2\|\tilde{\beta}^M - \beta_{(0)}^M\|_2 - \|\beta_{(0)}^M\|\max_{g\in K_0^c}\|\hat{\beta}_g - \beta_g\|_2 \\ & = \min_{g\in K_0^c}\beta_g^T\beta_{(0)}^M + o_p(1). \end{aligned} \tag{47}$$

Similarly, we can show

$$\max_{g \in K_0} \hat{\beta}_g^T \tilde{\beta}^M \le \max_{g \in K_0} \beta_g^T \beta_{(0)}^M + o_p(1). \tag{48}$$

It follows from lemma 3 that $\max_{g \in K_0} \beta_g^T \beta_{(0)}^M < \min_{g \in K_0^c} \beta_g^T \beta_{(0)}^M$. This together with (47) and (48) suggests that with probability tending to 1, we have

$$\min_{g \in K_0^c} \hat{\beta}_g^T \tilde{\beta}^M > \max_{g \in K_0} \beta_g^T \tilde{\beta}^M. \tag{49}$$

Step 4: It follows from lemma 3 that each element in the vector $\omega_0 = (\omega_0^{(1)}, \ldots, \omega_0^{(|K_0|)}) = (B_{K_0}^T B_{K_0})^{-1} e$ is nonnegative. Under Condition (C2), it is further assumed that all elements in ω_0 are nonzero. Hence, we obtain $\omega_0^{(k)} > 0$, $\forall k = 1, \ldots, |K_0|$.

Besides, it follows from (43) that

$$\|(\hat{B}_{K_0}^T \hat{B}_{K_0})^{-1} e - (B_{K_0}^T B_{K_0})^{-1} e\|_2 \le \sqrt{|K_0|} \|(\hat{B}_{K_0}^T \hat{B}_{K_0})^{-1} - (B_{K_0}^T B_{K_0})^{-1}\|_2 \xrightarrow{P} 0.$$

This implies that with probability tending to 1, all the elements in the vector $(\hat{B}_{K_0}^T \hat{B}_{K_0})^{-1} e$ are nonnegative. Combining results shown in Steps 2 and 3, it follows from lemma 3 that $\tilde{\beta}^M$ is the solution to the problem (12).

Step 5: As for $\hat{c}^M$, we have

$$|\hat{c}^M - c_{(0)}^M| = |\frac{\hat{c}_0}{\hat{G}_0} - \frac{c_0}{G_0}| \le \frac{1}{\hat{G}_0} |\hat{c}_0 - c_0| + |\frac{\min_g \beta_g^T \beta^M - \min_g \hat{\beta}_g^T \hat{\beta}^M}{\hat{G}_0 G_0}|.$$

By the definition of K_0, we have $\min_g \beta_g^T \beta_{(0)}^M = \min_{g \in K_0} \beta_g^T \beta_{(0)}^M$. When $\hat{\beta}^M = \tilde{\beta}^M$, it follows from (49) that $\min_g \hat{\beta}_g^T \hat{\beta}^M = \min_{g \in K_0} \hat{\beta}_g^T \hat{\beta}^M$. Hence, we obtain

$$\begin{aligned} & |\min_g \beta_g^T \beta_{(0)}^M - \min_g \hat{\beta}_g^T \hat{\beta}^M| = |\min_{g \in K_0} \beta_g^T \beta_{(0)}^M - \min_{g \in K_0} \hat{\beta}_g^T \hat{\beta}^M| \\ \le & \max_{g \in K_0} \|\beta_g\|_2 \|\hat{\beta}^M - \beta_{(0)}^M\|_2 + \|\hat{\beta}^M\|_2 \max_{g \in K_0} \|\hat{\beta}_g - \beta_g\|_2, \end{aligned}$$

with probability tending to 1. Convergence rate of $\hat{c}^M$ thus follows from those of $\hat{\beta}^M$ and $\hat{\beta}_g, g \in K_0$. This completes the proof.

E.6. Proof of theorem 6

In the proof of theorem 5, we have shown that

$$\begin{aligned} \hat{\beta}^M &= [e^T (\hat{B}_{K_0}^T \hat{B}_{K_0})^{-1} e]^{-1/2} \hat{B}_{K_0} (\hat{B}_{K_0}^T \hat{B}_{K_0})^{-1} e, \\ \beta_{(0)}^M &= [e^T (B_{K_0}^T B_{K_0})^{-1} e]^{-1/2} B_{K_0} (B_{K_0}^T B_{K_0})^{-1} e. \end{aligned}$$

Define $t_0 = B_{K_0}^T(B_{K_0}^T B_{K_0})^{-1}e$ and let e_g be the base vector with the gth coordinate as 1. For any matrix Ψ, denote by $N(\Psi)$ the projection matrix $I - \Psi(\Psi^T\Psi)^{+}\Psi^T$. For any vector $a \in \mathbb{R}^s$, it follows from the proof for theorem 5 that $a^T(\hat{\beta}^M - \beta^M)$ can be decomposed as $\eta_1 + \eta_2 + \eta_3$ where

$$\begin{aligned}
\eta_1 &= \left([e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e]^{-1/2} - \|t_0\|_2^{-1/2}\right) a^T t_0, \\
\eta_2 &= \frac{1}{\|t_0\|_2} a^T \left(\hat{B}_{K_0}(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1} - B_{K_0}(B_{K_0}^T B_{K_0})^{-1}\right) e, \\
\eta_3 &= \left([e^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e]^{-1/2} - t_0^{-1/2}\right)\left(a^T\hat{B}_{K_0}(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e - a^T t_0\right),
\end{aligned}$$

for any $a \in \mathbb{R}^s$ with $\|a\|_2 = 1$.

Condition (C3) implies all $\hat{\beta}_g$, $g \in K_0$ and $\hat{c}_0$ are $m^{-1/2}$ consistent. Using a second order Taylor expansion, we can show

$$\sqrt{m}\eta_1 = \sqrt{m}\sum_{g\in K_0}\frac{1}{\|t_0\|_2}\left(a^T N(B_{K_0})(\hat{\beta}_g - \beta_g)t_0^T v_g - a^T v_g t_0^T(\hat{\beta}_g - \beta_g)\right) + o_p(1),$$

and

$$\sqrt{m}\eta_2 = \frac{\sqrt{m}}{\|t_0\|^3}\sum_{g\in K_0} e^T(B_{K_0}^T B_{K_0})^{-1}e_g a^T t_0 t_0^T(\hat{\beta}_g - \beta_g) + o_p(1),$$

where $v_g = B_{K_0}(B_{K_0}^T B_{K_0})^{-1}e_g$, e_g is the basis vector with the gth coordinate equal to 1.

Using similar arguments in the proof of theorem 5, we can show

$$\left(e_K^T(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e_K\right)^{-1/2} - \left(e_K^T(B_{K_0}^T B_{K_0})^{-1}e_K\right)^{-1/2} = o_p(1),$$

and hence

$$\sqrt{m}\eta_3 = o_p(1)\sqrt{m}\left(a^T\hat{B}_{K_0}(\hat{B}_{K_0}^T\hat{B}_{K_0})^{-1}e - a^T t_0\right) = o_p(1)O_p(1) = o_p(1).$$

Combining these results together, we have that $\sqrt{m}(\hat{\beta}^M - \beta_{(0)}^M)$ is equivalent to

$$\frac{\sqrt{m}}{\|t_0\|_2}\sum_{g\in K_0}\left\{v_g^T e N(B_{K_0}) - e^T v_g N(t_0)\right\}(\hat{\beta}_g - \beta_g). \tag{50}$$

Under Condition (C3), the set of vectors $\{\sqrt{m}(\hat{\beta}_g - \beta_g), g \in K_0\}$ are asymptotically normally distributed. The asymptotic normality of $\hat{\beta}^M$ thus follows.

Similarly, we can show $\sqrt{m}(\hat{c}^M - c_{(0)}^M)$ is equivalent to

$$\sum_{g\in K_0}\sqrt{m}(\hat{\beta}_g - \beta_g)^T\psi_g + \psi_0\sqrt{m}(\hat{c}_0 - c_0), \tag{51}$$

for some constant ψ_0 and vectors $\psi_g, g \in K_0$. Hence, $\sqrt{m}(\hat{c}^M - c_{(0)}^M)$ is also asymptotically normal.

To derive the form of V^M, let $\Psi = (\psi_1, \dots, \psi_G)$ and Ψ_{K_0} be the matrix formed by columns in Ψ. Similarly define $U = (v_1, \dots, v_G)$ and U_{K_0}. For any $m \times n$ matrix A, denoted by vec(A) the

$mn \times 1$ vector obtained by stacking the columns of the matrix A on top of one another. For any vector $a = (a_1, \ldots, a_p)^T$ and matrix A, define the Kronecker product

$$a \otimes A = (a_1 A^T, \ldots, a_p A^T)^T.$$

Under Condition (C3), denoted by Φ the asymptotic covariance matrix of

$$Z = \begin{pmatrix} \sqrt{m}(\hat{c}_0 - c_{(0)}) \\ \sqrt{m}(\mathrm{vec}(\hat{B}_{K_0}) - \mathrm{vec}(B_{K_0})) \end{pmatrix}.$$

Besides, let

$$J = \frac{1}{\|t_0\|_2}\left[(U_{K_0}^T e) \otimes \{N(B_{K_0})^T - N(t_0)^T\}\right],$$

and $\Psi_0 = (\psi_0, \mathrm{vec}^T(\Psi))^T$. Using some algebra, it follows from (50) and (51) that

$$V^M = \lim_n \mathrm{cov} \begin{pmatrix} \sqrt{m}(\hat{c}^M - c^M_{(0)}) \\ \sqrt{m}(\hat{\beta}^M - \beta^M_{(0)}) \end{pmatrix} = \lim_n \mathrm{cov} \begin{pmatrix} \Psi_0^T Z \\ J^T Z \end{pmatrix} = (\Psi_0, J)^T \Phi (\Psi_0, J).$$

The proof is hence completed.

F. Technical lemmas and definitions

Lemma F.1. *For any vectors a, b subject to the constraint $\|a\|_2 = \|b\|_2 = 1$, there exits an orthogonal matrix Γ such that*

$$\Gamma a = (1, 0, \ldots, 0)^T, \quad \Gamma b = (c, \sqrt{1-c^2}, 0, \ldots, 0)^T,$$

where $c = a^T b$.

Proof: Denote Γ_1 to an arbitrary orthogonal matrix whose first column vector is equal to a. Assume $\Gamma_1 = (a, u_2, \ldots, u_p)^T$ where u_i is orthogonal to a for $i \geq 2$, we have

$$\Gamma_1 a = (1, 0, \ldots, 0)^T, \quad \Gamma_1 b = (c, d^T)^T,$$

where $d = (u_2^T b, \ldots, u_p^T b)$. Now define $\tilde{\Gamma}_2$ to be any $(p-1) \times (p-1)$ orthogonal matrix with the first column equal to $d/\|d\|_2$ and

$$\Gamma_2 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{\Gamma}_2 \end{pmatrix},$$

the assertion follows by setting $\Gamma = \Gamma_2 \Gamma_1$.

Definition F.1. *A $s \times 1$ random vector X is said to have a spherically symmetric distribution if for every $s \times s$ orthogonal matrix Γ, $\Gamma X \overset{d}{=} X$.*

Lemma F.2 (Fang et al. (1990)). *Assume $s \times 1$ random vector $X = (X_1, \ldots, X_s)^T$ is spherically distributed, then*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \overset{d}{=} \begin{pmatrix} rdU_1 \\ rdU_2 \end{pmatrix},$$

where U_1 and U_2 uniformly distributed on the L_2 ball: $u_1^2 + u_2^2 = 1$. $r \overset{d}{=} \|X\|_2 \geq 0$ and $d > 0$ are random variables distributed independently of U_1 and U_2.

Lemma F.3. *For any symmetric matrix A, we have $\|A\|_2 \leq \|A\|_\infty$.*

Proof: $\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty$. Since A is symmetric, $\|A\|_1 = \|A\|_\infty$ and the assertion follows.

Proof of lemma 4: Let $\beta^M = \arg\max_{\beta:\|\beta\|_2=1} \min_g \beta^T \beta_g$, for any β subject to the constraint $\|\beta\|_2 = 1$, there exists some j such that

$$\beta_j^T \beta \leq \min_g \beta_g^T \beta^M.$$

This together with the assumption on h suggests for any t, we have

$$h(\beta, \beta_j, t) \leq \min_g h(\beta^M, \beta_g, t),$$

which further implies

$$\mathrm{E}h(\beta, \beta_j, T) \leq \mathrm{E}\min_g h(\beta^M, \beta_g, T) \leq \min_g \mathrm{E}h(\beta^M, \beta_g, T).$$

The proof is hence completed.

Proof for lemma E.1, E.2 and E.3: Proofs for these lemmas are similar to those of lemma 2-4 in the paper Avi-Itzhak et al. (1995). We note that although in Avi-Itzhak et al. (1995), they require $\|\beta_g\|_2 = 1$ for all g, and β_g's to be linearly independent, these assumptions can be relaxed. We provide proofs for these lemmas below.

Proof for lemma E.1: If $K(\beta_0) = [1, \ldots, G]$, the proof becomes trivial. Otherwise, denoted by $S(\beta_0) = [1, \ldots, G] - K(\beta_0)$, we have $S(\beta_0) \neq \emptyset$. Let

$$\Delta_0 = \min_{g \in S(\beta_0)} \beta_0^T \beta_g - F(\beta_0).$$

By definition, we have $\Delta_0 > 0$.

Let $M_0 = \max_{g=1,\ldots,G} \|\beta_g\|_2$, $\varepsilon = \Delta_0/(3M_0)$. For any β such that $\|\beta - \beta_0\|_2 < \varepsilon$, we have

$$|\beta^T \beta_g - \beta_0^T \beta_g| < M_0 \varepsilon = \frac{\Delta_0}{3},$$

and hence

$$\beta_0^T\beta_g - \frac{\Delta_0}{3} < \beta^T\beta_g < \beta_0^T\beta_g + \frac{\Delta_0}{3}. \tag{52}$$

The first inequality in (52) implies

$$\min_{g\in S(\beta_0)} \beta^T\beta_g > \min_{g\in S(\beta_0)} \beta_0^T\beta_g - \Delta_0/3, \tag{53}$$

while the second inequality in (52) suggests

$$\max_{g\in K(\beta_0)} \beta^T\beta_g < \max_{g\in K(\beta_0)} \beta_0^T\beta_g + \Delta_0/3 = F_0 + \Delta_0/3. \tag{54}$$

Combining (53) with (54), we obtain

$$\max_{g\in K(\beta_0)} \beta^T\beta_g < F_0 + \frac{\Delta_0}{3} \le \min_{g\in S(\beta_0)} \beta_0^T\beta_g - \Delta_0 + \frac{\Delta_0}{3} < \min_{g\in S(\beta_0)} \beta^T\beta_g.$$

The proof is hence completed.

Proof for lemma E.2: Define K_0 by

$$K_0 = \{g \in \{1,\dots,G\} : \beta_g^T\beta^M = \min_j \beta_j^T\beta^M\}.$$

Obviously, K_0 is unique. Besides, by definition, we have

$$\min_{g\in K_0^c} \beta_g^T\beta^M > \beta_j^T\beta^M, \quad \forall j \in K_0.$$

Moreover, since $\beta_j^T\beta^M$ are the same for all $j \in K_0$, we have $\beta^M \in \mathrm{E}_{K_0}(B)$.

Proof for lemma E.3: Any unit length vector β_0 can be represented as

$$\beta_0 = \frac{\beta+\delta}{\|\beta+\delta\|_2},$$

for some vector δ. We further decompose $\delta = \delta_1 + \delta_2$ where $\delta_1 \in C(B_{K_0})$ and $\delta_2^T B_{K_0} = 0$. Let

$$\beta' = \frac{\beta+\delta_1}{\|\beta+\delta_1\|_2}.$$

Since $\beta \in C(B_{K_0})$, we have $\beta' \in C(B_{K_0})$ and $\delta_2^T\beta = 0$.

Observe that

$$\begin{aligned}\|\beta' - \beta\|_2^2 &= 2 - 2\beta^T\beta' = 2 - 2\frac{1+\beta^T\delta_1}{\sqrt{1+\|\delta_1\|_2^2 + 2\beta^T\delta_1}} \\ &\le 2 - 2\frac{1+\beta^T\delta_1}{\sqrt{1+\|\delta\|_2^2 + 2\beta^T\delta_1}} = \|\beta_0 - \beta\|_2^2 < \varepsilon^2.\end{aligned}$$

This implies β' also lies within the ε-neighborhood of β. Besides, for any $g \in K_0$,

$$\beta_g^T\beta_0 = \frac{\beta_g^T(\beta+\delta)}{\sqrt{1+\|\delta\|_2^2+2\beta^T\delta}} \le \frac{\beta_g^T(\beta+\delta_1)}{\sqrt{1+\|\delta_1\|_2^2+2\beta^T\delta_1}} = \beta_g^T\beta'.$$

Therefore, we obtain $F_{K_0}(\beta') \ge F_{K_0}(\beta_0)$. The proof is hence completed.

References

Avi-Itzhak, H., J. A. Van Mieghem, L. Rub, et al. (1995). Multiple subclass pattern recognition: a maximin correlation approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 17*(4), 418–431.

Fang, K. T., S. Kotz, and K. W. Ng (1990). *Symmetric multivariate and related distributions*, Volume 36 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, Ltd., London.

Kay, S. R., A. Fiszbein, and L. A. Opfer (1987). The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia bulletin 13*(2), 261.

Tarrier, N., S. Lewis, G. Haddock, R. Bentall, R. Drake, P. Kinderman, D. Kingdon, R. Siddle, J. Everitt, K. Leadley, et al. (2004). Cognitive-behavioural therapy in first-episode and early schizophrenia. *The British Journal of Psychiatry 184*(3), 231–239.