Testing for treatment effect in multitreatment case

Pier Luigi Conti, Livia De Giovanni, Ayoub Mounim

Università di Roma “La Sapienza”, Piazzale Aldo Moro 5, Rome, Italy; Luiss University, Viale Romania 32, Rome, Italy

Supplementary Materials

S1. Estimation of generalized propensity scores

In this section the estimation of the generalized propensity scores $p_k(\boldsymbol{x})$ is shortly discussed. First of all, as a consequence of Lorenz (1986) (Th. 8, p. 80), if the “true” propensity scores $p_k(\boldsymbol{x})$ are s times continuously differentiable, then there exist L-dimensional vectors $\boldsymbol{\pi}_{k,L}$, $k = 1, \ldots, K$ such that, for a sequence of orthonormal polynomials $\boldsymbol{x}_{vec,L}$, such that

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left| p_k(\boldsymbol{x}) - \frac{\exp\{\boldsymbol{x}_{vec,L}^T \boldsymbol{\pi}_{k,L}\}}{1+\sum_{k=1}^K \exp\{\boldsymbol{x}_{vec,L}^T \boldsymbol{\pi}_{k,L}\}} \right| = O\left(L^{-s/P}\right)$$

$\boldsymbol{x}_{vec,L}$ being a L-dimensional vector of coefficients. The orthogonal polynomials in $\boldsymbol{x}_{vec,L}$ are explicitly constructed in Hirano et al. (2003).

The above considerations would naturally suggest to use a nonparametric logistic polytomous model, by adopting

$$\begin{aligned} p_0^w(\boldsymbol{x}) &= \frac{1}{1+\sum_{k=1}^K \exp\{\boldsymbol{x}_{vec,L}^T \boldsymbol{\pi}_{k,L}\}}, \\ p_k^w(\boldsymbol{x}) &= \frac{\exp\{\boldsymbol{x}_{vec,L}^T \boldsymbol{\pi}_{k,L}\}}{1+\sum_{k=1}^K \exp\{\boldsymbol{x}_{vec,L}^T \boldsymbol{\pi}_{k,L}\}}, \quad k = 1, \ldots, K \end{aligned} \tag{S1.1}$$

as a working model for $p_k(\boldsymbol{x})$.

The vectors $\boldsymbol{\pi}_{k,L}$ are then estimated through sieve maximum likelihood estimators. Consider first the functions

$$h_k(\boldsymbol{x}) = \log \frac{p_k(\boldsymbol{x})}{1 + \sum_{k=1}^{K} p_k(\boldsymbol{x})},\ k = 1, \ldots, K. \tag{S1.2}$$

that They are assumed to live in the Hölder space H^q composed by all functions $g:\ \mathcal{X} \to \mathbb{R}$ possessing first $[q]$ derivatives that are bounded, and the $[q]$th derivatives are Hölder continuous of order $q - [q]$, where $[q]$ is the integer part of q. Using the orthogonal polynomials in $\boldsymbol{x}_{vec,L}$, a sequence of finite-dimensional sieve spaces $H_L = \left\{h(\boldsymbol{x}) = \boldsymbol{x}_{vec,L}^T \boldsymbol{\pi}_{k,L};\ \|h\|_q a \leq c\right\}$ is constructed, that approximate the functions in H^δ in terms of the Soboles norm, defined as

$$\|g\|_q^S = \max_{a_1+\cdots+a_P \leq q} \sup_{\boldsymbol{x}\in\mathcal{X}} |\nabla^a g(\boldsymbol{x})| + \max_{a_1+\cdots+a_P \leq q} \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{X}} \frac{|\nabla^a g(\boldsymbol{x}) - \nabla^a g(\boldsymbol{y})|}{(\|\boldsymbol{x} - \boldsymbol{y}\|_2)^{q-[q]}}$$

where $\|\cdot\|_2$ is the Euclidean norm, a is the vector of components $a_1 \geq 0, \ldots, a_P \geq 0$, and

$$\nabla^a g(\boldsymbol{x}) = \frac{\partial^{a_1+\cdots+a_P}}{\partial x_1^{a_1} \cdots \partial x_1^{a_1}} g(\boldsymbol{x}). \tag{S1.3}$$

The vectors $\boldsymbol{\pi}_{k,L}$ are then estimated by maximizing the working log-likelihood

$$\widehat{\boldsymbol{\pi}}_{k,L} = \arg\max \sum_{k=0}^{K} \sum_{\substack{i=1 \\ T_i=k}}^{n} \log p_k^w(\boldsymbol{x}_i),\quad k = 1, \ldots, K.$$

As estimators of $p_k(\boldsymbol{x})$s, it is then natural to take

$$\begin{aligned} \widehat{p}_0(\boldsymbol{x}) &= \frac{1}{1 + \sum_{k=1}^{K} \exp\{\boldsymbol{x}_{vec,L}^T \widehat{\boldsymbol{\pi}}_{k,L}\}}, \\ \widehat{p}_k(\boldsymbol{x}) &= \frac{\exp\{\boldsymbol{x}_{vec,L}^T \widehat{\boldsymbol{\pi}}_{k,L}\}}{1 + \sum_{k=1}^{K} \exp\{\boldsymbol{x}_{vec,L}^T \widehat{\boldsymbol{\pi}}_{k,L}\}}, k = 1, \ldots, K. \end{aligned}$$

As a minor generalization of results in Hirano et al. (2003) (cfr. also Kim (2013)), the following result is obtained.

Proposition 1. *Suppose that assumptions H1-H3 are verified, and that*

H4. The functions $h_k(\boldsymbol{x}) = \log \frac{p_k(\boldsymbol{x})}{1+\sum_{k=1}^{K} p_k(\boldsymbol{x})}$, $k = 1, \dots, K$ are $s > P$ times differentiable, and possess finite Sobolev norm of order $q > P/2$.

If $L = L_n = Cn^t$ with $\frac{1}{4(s/P-1)} < t < 1/6$, then

$$\sup_{\boldsymbol{x} \in \mathcal{X}} |\widehat{p}_k(\boldsymbol{x}) - p_x(\boldsymbol{x})| = o_p(n^{-1/4}), \ k = 1, \dots, K. \tag{S1.4}$$

S2. Proofs of main results

Proof of Proposition 3. Define first, as in Proposition , $W_{k,n}(y) = \sqrt{n}(\widehat{F}_k(y) - F_k(y))$. By an integration by parts, the relationships

$$\begin{aligned}
\sqrt{n}(\widehat{\theta}_{jk} - \theta_{jk}) &= \sqrt{n}\int_{-\infty}^{+\infty} (\widehat{F}_j(y) - F_j(y))\, d[\widehat{F}_k(y) - F_k(y)] \\
&\quad + \sqrt{n}\int_{-\infty}^{+\infty} F_j(y)\, d[\widehat{F}_k(y) - F_k(y)] + \sqrt{n}\int_{-\infty}^{+\infty} (\widehat{F}_j(y) - F_j(y))\, dF_k(y) \\
&= \sqrt{n}\int_{-\infty}^{+\infty} (\widehat{F}_j(y) - F_j(y))\, d[\widehat{F}_k(y) - F_k(y)] \\
&\quad + \sqrt{n}\, F_j(y)(\widehat{F}_k(y) - F_k(y))\Big|_{-\infty}^{+\infty} - \sqrt{n}\int_{-\infty}^{+\infty} (\widehat{F}_k(y) - F_k(y))\, dF_j(y) \\
&\quad + \sqrt{n}\int_{-\infty}^{+\infty} (\widehat{F}_j(y) - F_j(y))\, dF_k(y) \\
&= \int_{-\infty}^{+\infty} W_{j,n}(y)\, d\left[\frac{1}{\sqrt{n}} W_{k,n}(y)\right] + \int_{-\infty}^{+\infty} W_{j,n}(y)\, dF_k(y) \\
&\quad - \int_{-\infty}^{+\infty} W_{k,n}(y)\, dF_j(y)
\end{aligned} \tag{S2.5}$$

are obtained, for every $j \neq k = 0, 1, \dots, K$. Next, form the Skorokhod representation theorem (Billingsley (1999), p. 70), there exists on an appropriate probability space

$(\widetilde{\Omega},\, \widetilde{\mathcal{F}},\, \widetilde{P})$, vector stochastic processes

$$\widetilde{\boldsymbol{W}}_n(y) = \begin{bmatrix} \widetilde{\boldsymbol{W}}_{0\,n}(y) \\ \widetilde{\boldsymbol{W}}_{1\,n}(y) \\ \cdots \\ \widetilde{\boldsymbol{W}}_{K\,n}(y) \end{bmatrix}, \quad \widetilde{\boldsymbol{W}}(y) = \begin{bmatrix} \widetilde{\boldsymbol{W}}_{0}(y) \\ \widetilde{\boldsymbol{W}}_{1}(y) \\ \cdots \\ \widetilde{\boldsymbol{W}}_{K}(y) \end{bmatrix}$$

such that

(i) $\widetilde{\boldsymbol{W}}_n(y) \overset{d}{=} \boldsymbol{W}_n(y)\ \forall\, n \geq 1$;

(ii) $\widetilde{\boldsymbol{W}}(y) \overset{d}{=} \boldsymbol{W}(y)$;

(iii) $\max_{0 \leq k \leq K} \sup_y \left| \widetilde{\boldsymbol{W}}_n(y) - \widetilde{\boldsymbol{W}}(y) \right| \to 0$ with $\widetilde{P}$-probability 1 as $n \to \infty$.

Hence, we may write

$$\sqrt{n}(\widehat{\theta}_{jk} - \theta_{jk}) \overset{d}{=} V^{jk}_{1,n} + V^{jk}_{2,n} - V^{jk}_{3,n} \tag{S2.6}$$

where

$$\begin{aligned} V^{jk}_{1,n} &= \int_{-\infty}^{+\infty} \widetilde{W}_{j,n}(y)\, d\left[\frac{1}{\sqrt{n}} \widetilde{W}_{k,n}(y)\right], \\ V^{jk}_{2,n} &= \int_{-\infty}^{+\infty} \widetilde{W}_{j,n}(y)\, dF_k(y), \\ V^{jk}_{3,n} &= -\int_{-\infty}^{+\infty} \widetilde{W}_{k,n}(y)\, dF_j(y). \end{aligned}$$

As far as the term $V^{jk}_{1,n}$ above is concerned, we have

$$\int_{-\infty}^{+\infty} \widetilde{W}_{j,n}(y)\, dF_k(y) \;=\; \int_{-\infty}^{+\infty} \widetilde{W}_{j}(y)\, dF_k(y) + \int_{-\infty}^{+\infty} (\widetilde{W}_{j,n}(y) - \widetilde{W}_{j}(y))\, dF_k(y)$$

with

$$\left| \int_{-\infty}^{+\infty} (\widetilde{W}_{j,n}(y) - \widetilde{W}_j(y))\, dF_k(y) \right| \leq \sup_y \left| \widetilde{W}_{j,n}(y) - \widetilde{W}_j(y) \right| \to 0 \;\; a.s. - \widetilde{P}$$

and hence, as $n \to \infty$,

$$V_{2,n}^{jk} \to V_2^{jk} = \int_{-\infty}^{+\infty} \widetilde{W}_j(y)\, dF_k(y) \;\; a.s. - \widetilde{P}. \tag{S2.7}$$

In the same way, it can be shown that, as $n \to \infty$,

$$V_{3,n}^{jk} \to V_3^{jk} = \int_{-\infty}^{+\infty} \widetilde{W}_k(y)\, dF_j(y) \;\; a.s. - \widetilde{P}. \tag{S2.8}$$

Finally, as far as the term V_{1n}^{jk} is concerned, observe first that the signed measure corresponding to $\widetilde{W}_{k,n}(y)/\sqrt{n}$ has total variation not larger than 2, and hence

$$\begin{aligned} \left| V_{3,n}^{jk} \right| &\leq \left| \int_{-\infty}^{+\infty} \widetilde{W}_j(y)\, d \left[\frac{1}{\sqrt{n}} \widetilde{W}_{k,n}(y) \right] \right| + \left| \int_{-\infty}^{+\infty} (\widetilde{W}_{j,n}(y) - \widetilde{W}_j(y))\, d \left[\frac{1}{\sqrt{n}} \widetilde{W}_{k,n}(y) \right] \right| \\ &\leq \left| \int_{-\infty}^{+\infty} \widetilde{W}_j(y)\, d \left[\frac{1}{\sqrt{n}} \widetilde{W}_{k,n}(y) \right] \right| + 2 \sup_y \left| \widetilde{W}_{j,n}(y) - \widetilde{W}_j(y) \right| \end{aligned}$$

with

$$\left| \int_{-\infty}^{+\infty} \widetilde{W}_j(y)\, d \left[\frac{1}{\sqrt{n}} \widetilde{W}_{k,n}(y) \right] \right| \to 0 \;\; a.s. - \widetilde{P}$$

because of the Helly-Bray theorem, and $\sup_y \left| \widetilde{W}_{j,n}(y) - \widetilde{W}_j(y) \right|$ tending to 0 a.s.-$\widetilde{P}$ because of (iii). This shows that $A_{3,n} \to 0$ with $\widetilde{P}$-probability 1, and hence

$$V_{1,n}^{jk} + V_{2,n}^{jk} - V_{3,n}^{jk} \to V_2^{jk} - V_3^{jk} = \int_{-\infty}^{+\infty} \widetilde{W}_j(y)\, dF_k(y) - \int_{-\infty}^{+\infty} \widetilde{W}_k(y)\, dF_j(y) \tag{S2.9}$$

$a.s. - \widetilde{P}$, as $n \to \infty$.

Define now the random vectors

$$\boldsymbol{V}_n = \begin{bmatrix} V_{1,n}^{01} + V_{2,n}^{01} - V_{3,n}^{01} \\ V_{1,n}^{02} + V_{2,n}^{02} - V_{3,n}^{02} \\ \cdots \\ V_{1,n}^{0K} + V_{2,n}^{0K} - V_{3,n}^{0K} \\ V_{1,n}^{12} + V_{2,n}^{13} - V_{3,n}^{1K} \\ V_{1,n}^{13} + V_{2,n}^{13} - V_{3,n}^{13} \\ \cdots \\ V_{1,n}^{1K} + V_{2,n}^{1K} - V_{3,n}^{1K} \\ \cdots \\ V_{1,n}^{K-1\,K} + V_{2,n}^{K-1\,K} - V_{3,n}^{K-1\,K} \end{bmatrix}, \quad \boldsymbol{V} = \begin{bmatrix} V_2^{01} - V_3^{01} \\ V_2^{02} - V_3^{02} \\ \cdots \\ V_2^{0K} - V_3^{0K} \\ V_2^{13} - V_3^{1K} \\ V_2^{13} - V_3^{13} \\ \cdots \\ V_2^{1K} - V_3^{1K} \\ \cdots \\ V_2^{K-1\,K} - V_3^{K-1\,K} \end{bmatrix}. \tag{S2.10}$$

Eqns. $(S2.9)$, being true for each j, k, imply that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{vec} - \boldsymbol{\theta}_{vec}) \overset{d}{=} \boldsymbol{V}_n \to \boldsymbol{V} \;\; a.s. - \widetilde{P}$$

as $n \to \infty$, which implies, in its turn

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{vec} - \boldsymbol{\theta}_{vec}) \xrightarrow{d} \boldsymbol{V} \;\; \text{as } n \to \infty. \tag{S2.11}$$

Finally, the map $\boldsymbol{W} \mapsto \boldsymbol{V}$ is a linear map of a (centered) Gaussian process, and hence it possesses Multivariate Normal distribution with null mean vector and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{V}}$. As far as the identification of $\boldsymbol{\Sigma}_{\boldsymbol{V}}$ is concerned, its elements are essentially the asymptotic covariances of $\widehat{\theta}_{jk}$ and $\widehat{\theta}_{hl}$ (when $h = j$ and $l = k$ they reduce to asymptotic variances). They can be written as

$$
\begin{aligned}
\sigma_{jk;hl} &= E\left[\left(\int_{-\infty}^{+\infty} W_j(y)\, dF_k(y) - \int_{-\infty}^{+\infty} W_k(y)\, dF_j(y)\right)\right. \\
&\quad \left.\left(\int_{-\infty}^{+\infty} W_h(y)\, dF_l(y) - \int_{-\infty}^{+\infty} W_l(y)\, dF_h(y)\right)\right] \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} E\left[W_j(y)W_h(t)\right]\, dF_k(y)\, dF_l(t) + \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} E\left[W_k(y)W_l(t)\right]\, dF_j(y)\, dF_h(t) \\
&- \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} E\left[W_j(y)W_l(t)\right]\, dF_k(y)\, dF_h(t) \\
&- \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} E\left[W_k(y)W_h(t)\right]\, dF_j(y)\, dF_l(t) \qquad \text{(S2.12)}
\end{aligned}
$$

and the covariances $C_{jh} = E\left[W_j(y)W_h(t)\right]$ are in (12). $\square$

Proof of Proposition 5. Define the matrices $\widehat{\boldsymbol{A}}$, $\widehat{\boldsymbol{\Lambda}}$ exactly ad $\boldsymbol{A}$, $\boldsymbol{\Lambda}$, but with λ_k replaced by n_k/n, $k = 0, 1, \dots, K$. The Strong Law of Large Numbers implies that $n_k/n \overset{a.s.}{\to} \lambda_k$ for all $k = 0, 1, \dots, K$, so that $\widehat{\boldsymbol{A}} \overset{a.s.}{\to} \boldsymbol{A}$, $\widehat{\boldsymbol{\Lambda}} \overset{a.s.}{\to} \boldsymbol{\Lambda}$, where a.s. convergence of matrices is component-wise. When $F_0 = F_1 = \cdots = F_K$, all components of the vector $\boldsymbol{\theta}_{vec}$ are equal to 1/2, so that

$$
D_n = n(\widehat{\boldsymbol{\theta}}_{vec} - \boldsymbol{\theta}_{vec})^T \widehat{\boldsymbol{A}}^T \widehat{\boldsymbol{\Lambda}} \widehat{\boldsymbol{A}} (\widehat{\boldsymbol{\theta}}_{vec} - \boldsymbol{\theta}_{vec})
$$

statement 1 follows from Proposition 4 and Slutsky Theorem.

As far as Statement 2 is concerned, it is enough to observe that, as a consequence of the Strong Law of Large Numbers, $D_n/n \overset{a.s.}{\to} \delta$ as n increases. If $\delta > 0$, eqn. (33) easily follows. $\square$

S3. Simulation Study 2: exact distribution of potential outcomes

In case $K + 1 = 3$, under Scenario I (zero treatment effect), the exact distribution function of $Y_{(k)}$ is

$$F_k(y) = \begin{cases} 0 & y < 60 \\ \frac{y-60}{40}\left(\frac{1}{2}\cdot\frac{y-60}{20}\right) & 60 \le y < 70 \\ \frac{y-65}{20}\left(\frac{1}{2}\cdot\frac{y-60}{20} + \frac{1}{2}\cdot\frac{y-70}{20}\right) & 70 \le y < 80 \\ \frac{y-50}{40}\left(\frac{1}{2} + \frac{1}{2}\cdot\frac{y-70}{20}\right) & 80 \le y < 90 \\ 1 & y \ge 90 \end{cases}, \quad k = 0,\, 1,\, 2.$$

Furthermore we have $E[Y_{(0)}|T = 0] = 72.0$, $E[Y_{(1)}|T = 1] = 75.0$ and $E[Y_{(2)}|T = 2] = 78.0$, so that $E[Y_{(1)}|T = 1] - E[Y_{(0)}|T = 0] = 3.0$, $E[Y_{(2)}|T = 2] - E[Y_{(0)}|T = 0] = 6.0$, $E[Y_{(2)}|T = 2] - E[Y_{(1)}|T = 1] = 3.0$. The confounding effect of X makes it difficult to detect the absence of treatment effect.

Under Scenario II (non-zero treatment effect), the exact distribution functions of potential outcomes are reported below.

$$F_0(y) = \begin{cases} 0 & y < 64 \\ \frac{y-64}{40} & 60 \le y < 104 \\ 1 & y \ge 104 \end{cases} \quad F_1(y) = \begin{cases} 0 & y < 65 \\ \frac{y-65}{40} & 65 \le y < 105 \\ 1 & y \ge 105 \end{cases} \quad F_2(y) = \begin{cases} 0 & y < 66 \\ \frac{y-66}{40} & 66 \le y < 106 \\ 1 & y \ge 110 \end{cases}$$

In this case, we have $\theta_{01} = 0.52$, $\theta_{02} = 0.55$, $\theta_{12} = 0.52$, $E[Y_{(0)}] = 84.0$, $E[Y_{(1)}] = 85.0$, $E[Y_{(2)}] = 86.0$. Furthermore it is easy to see that $E[Y_{(0)}|T = 0] = E[Y_{(1)}|T = 1] = E[Y_{(2)}|T = 2] = 85.0$, so that $E[Y_{(1)}|T = 1] - E[Y_{(0}|T = 0] = E[Y_{(2)}|T = 2] - E[Y_{(0)}|T = 0] = E[Y_{(2)}|T = 2] - E[Y_{(1)}|T = 1] = 0.0$. The confounding effect of X again makes it difficult to detect the presence of treatment effect.

S4. Simulation Study 2: Comparison of Kruskal-Wallis type test and matching *GPSM* test

Table 1: Rejection probabilities (nominal significance level 0.95) - $K+1$=3

Kruskal-Wallis	n=500	n=1000	n=1500
I (H_0 true)	0.10	0.08	0.06
II (H_1 true)	0.90	0.98	1.00
GPSM	n=500	n=1000	n=1500
I (H_0 true)	0.12	0.10	0.07
II (H_1 true)	0.79	0.94	1.00

Table 2: Rejection probabilities (nominal significance level 0.95) - $K+1$=4

Kruskal Wallis	n=500	n=1000	n=2000
I (H_0 true)	0.11	0.06	0.05
II (H_1 true)	0.92	1.00	1.00
GPSM	n=500	n=1000	n=2000
I (H_1 true)	0.12	0.07	0.05
II (H_1 true)	0.88	1.00	1.00

References

Billingsley, P. (1999). *Convergence of Probability Measures* 2nd Ed. New York: Wiley.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica 71*, 1161–1189.

Kim, K. (2013). An Alternative Efficient Estimation of Average Treatment Effects. *Journal of Market Economy 42*, 1–41.

Lorenz, G. G. (1986). *Approximation of Functions.* Chelsea, New York.