Electronic Journal of Statistics
Vol. 8 (2014) 1312–1321
ISSN: 1935-7524
DOI: 10.1214/14-EJS920REJ

Rejoinder of "Dynamic treatment regimes: Technical challenges and applications"*

Eric B. Laber[†]

North Carolina State University
Raleigh, NC 27696-8203
e-mail: laber@stat.ncsu.edu

Daniel J. Lizotte

University of Waterloo
Waterloo, Ontario N2L 3G1
e-mail: dlizotte@uwaterloo.ca

Min Qian

Columbia University
New York, NY 10032
e-mail: mq2158@columbia.edu

William E. Pelham

Florida International University
Miami, FL 33199
e-mail: wpelham@fiu.edu

and

Susan A. Murphy

University of Michigan
Ann Arbor, MI 48106-1248
e-mail: samurphy@umich.edu

Received November 2013.

[*] Main article 10.1214/14-EJS920.

[†] Eric B. Laber is in the Department of Statistics at North Carolina State University, 2311 Stinson Dr., Raleigh, NC, 27695 (E-mail *laber@stat.ncsu.edu*). He acknowledges support from NIH grant P01 CA142538 and PR Department of Natural and Environmental Resources (PRDNER) grant PR-W-F14AF00171. Daniel J. Lizotte. is in the Department of Computer Science at the University of Waterloo, Ontario, N2L G31. He acknowledges support from the Natural Sciences and Engineering Research Council of Canada. Min Qian is in the Department of Biostatistics at Columbia University, New York City, NY, 10032. She acknowledges support from NSF grant DMS-1307838. Susan A. Murphy is in the Departments of Statistics and Psychiatry and in the Institute for Social Research, all at the University of Michigan, Ann Arbor, MI, 48109. She acknowledges support from NIMH grant R01-MH-080015 and NIDA grant P50-DA-010075.

Contents

We thank the discussants for their thoughtful and provocative discussions; they have raised several fundamental philosophical questions, suggested new and interesting directions for research, and established new properties of our proposed confidence interval approach. We also thank the Editor for organizing these discussions and for providing helpful feedback on earlier versions of our manuscript. Because the discussants touched on similar themes, we have organized our responses into categories: (i) choosing an appropriate target for inference; (ii) new approaches to estimation and inference; and (iii) scalability and direct estimation.

1. Choosing an appropriate target for inference

1.1. Confidence intervals for $c^\intercal\beta_1^*$

Consider two potential applications of confidence intervals for $c^\intercal\beta_1^*$: (U1) forming a confidence interval for $Q_1(h_1, 1) - Q_1(h_1, -1)$ to determine if there is sufficient evidence to recommend one treatment over the other for a patient presenting at baseline with $H_1 = h_1$; and (U2) to construct confidence intervals for individual components of β_1^* to determine which patient characteristics are important for tailoring treatment. Robins and Rotnitzky (RR hereafter) argue that for (U1) β_1^* is not clinically meaningful as it assumes that the decision maker will select treatment using π_2^{dp} at the second stage, which is likely untrue. RR propose developing confidence intervals for $c^\intercal\beta_1^*(\widehat{\pi}_2)$ defined as follows. For any second stage decision rule, say π_2, define

$$\beta_1^*(\pi_2) = \arg\min_{\beta_1} P\left\{H_{2,0}^\intercal\beta_{2,0}^* + \pi_2(H_2)H_{2,1}^\intercal\beta_{2,1}^* - H_{1,0}^\intercal\beta_{1,0} - A_1H_{1,1}^\intercal\beta_{1,1}\right\}^2,$$

so that $\beta_1^*(\pi_2)$ denotes the optimal first stage coefficients assuming that a decision maker will assign treatments according to π_2 at the second stage. Thus, RR argue that if the estimated optimal regime $\widehat{\pi}_2$ is to be used to assign treatments at the second stage then the data-dependent parameter $\beta_1^*(\widehat{\pi}_2)$ may be more clinically meaningful than $\beta_1^*(= \beta_1^*(\pi_2^{\mathrm{dp}})$ in their notation). In a pleasant surprise for us, RR showed that our proposed adaptive confidence interval $c^\intercal\beta_1^*$ is also a valid confidence interval for $c^\intercal\beta_1^*(\widehat{\pi}_2)$ at least in the case of binary predictors. We look forward to generalizations of this result.

We maintain that if (U2) is of interest, that is, the confidence intervals are used to inform scientific theory and generate hypotheses for subsequent studies, then confidence sets for β_1^* are scientifically meaningful. In this context it may be of interest to identify which patient covariates are important for optimal treatment choice at each stage. One way to identify important covariates under the optimal regime is to look at confidence intervals for components of $\beta_{2,1}^*$ and $\beta_{1,1}^*$. In our experience, estimation of optimal DTRs is often conducted as a secondary, exploratory analysis in which case confidence intervals for β_1^* are likely of interest.

1.2. An alternative value function

Goldberg, Song, Zeng, and Kosorok (GSZK hereafter) propose a new measure of the quality of a DTR. Traditionally the value of a DTR, say π, is defined as the expected outcome if all patients in the population of interest are assigned treatment according to π (see Schulte et al., 2013). Because the value of an estimated optimal DTR is nonregular when a non-null subgroup of patients have a small treatment effect in one or more of the treatment stages, GSZK propose a new estimand, called the truncated value of π, which is the expected outcome under π but restricted to the population of patients with clinically meaningful treatment effects at each stage. It is claimed (p. 6) that the truncated value can be made to be arbitrarily close to the value yet is regular and asymptotically normal under regularity conditions. We show that if one can obtain a regular estimand whose distance from the value can be controlled then one can obtain valid confidence intervals for value. We also demonstrate in a one-stage setting that the proposed truncated value may be arbitrary far from the value under certain generative models.

For an estimated DTR, say $\widehat{\pi}_n$, let $V(\widehat{\pi}_n) = \mathbb{E}^{\widehat{\pi}_n}Y$ be the value of $\widehat{\pi}_n$. Suppose that there exists an alternative estimand $V_\epsilon(\widehat{\pi}_n)$ for which: (P1) there exists estimator $\widehat{V}_{\epsilon,n}(\widehat{\pi}_n)$ so that $\sqrt{n}(\widehat{V}_{\epsilon,n}(\widehat{\pi}_n) - V_\epsilon(\widehat{\pi}_n))$ is regular and asymptotically normal; and (P2) $|V_\epsilon(\widehat{\pi}_n) - V(\widehat{\pi}_n)| \le b(\epsilon) + o_P(1)$ for some (possibly random) function $b(\epsilon)$ for which there exists consistent estimator $\widehat{b}_n(\epsilon)$ that satisfies $P(|V_\epsilon(\widehat{\pi}_n) - V(\widehat{\pi}_n)| \le \widehat{b}_n(\epsilon)) = 1 - o(1)$. Because $\widehat{V}_{\epsilon,n}(\widehat{\pi}_n)$ is regular and asymptotically normal for $\alpha \in (0,1)$ we can apply standard methods to construct consistent estimators, say $\widehat{u}_n$ and $\widehat{\ell}_n$, of the $(1-\alpha/2)\times 100\%$ and $(\alpha/2)\times 100\%$ percentiles of the sampling distribution of $\sqrt{n}(\widehat{V}_{\epsilon,n}(\widehat{\pi}_n) - V_\epsilon(\widehat{\pi}_n))$. Then $[\widehat{V}_{\epsilon,n}(\widehat{\pi}_n) - \widehat{u}_n/\sqrt{n} - \widehat{b}(\epsilon), \widehat{V}_{\epsilon,n}(\widehat{\pi}_n) - \widehat{\ell}_n/\sqrt{n} + \widehat{b}(\epsilon)]$ is a valid asymptotic $(1-\alpha)\times 100\%$ confidence interval for $V(\widehat{\pi}_n)$ because:

$$P\left(\widehat{V}_{\epsilon,n}(\widehat{\pi}_n) - \widehat{u}_n/\sqrt{n} - \widehat{b}(\epsilon) \le V(\widehat{\pi}_n) \le \widehat{V}_{\epsilon,n}(\widehat{\pi}_n) - \widehat{\ell}_n/\sqrt{n} + \widehat{b}(\epsilon)\right)$$

$$
\begin{aligned}
&\geq P\left(\widehat{V}_{\epsilon,n}(\widehat{\pi}_n) - \widehat{u}_n/\sqrt{n} - \widehat{b}(\epsilon) \leq V(\widehat{\pi}_n) \leq \widehat{V}_{\epsilon,n}(\widehat{\pi}_n) - \widehat{\ell}_n/\sqrt{n} + \widehat{b}(\epsilon),\right.\\
&\qquad\qquad\qquad\qquad \left. |V_\epsilon(\widehat{\pi}_n) - V(\widehat{\pi})| \leq \widehat{b}(\epsilon)\right)\\
&\qquad \geq P\left(\widehat{V}_{\epsilon,n}(\widehat{\pi}_n) - \widehat{u}_n/\sqrt{n} \leq V_\epsilon(\widehat{\pi}_n) \leq \widehat{V}_{\epsilon,n}(\widehat{\pi}_n) - \widehat{\ell}_n/\sqrt{n},\right.\\
&\qquad\qquad\qquad\qquad \left. |V_\epsilon(\widehat{\pi}_n) - V(\widehat{\pi})| \leq \widehat{b}(\epsilon)\right)\\
&= P\left(\widehat{\ell}_n \leq \sqrt{n}(\widehat{V}_{\epsilon,n}(\widehat{\pi}_n) - V_\epsilon(\widehat{\pi}_n)) \leq \widehat{u}_n,\, |V_\epsilon(\widehat{\pi}_n) - V(\widehat{\pi})| \leq \widehat{b}(\epsilon)\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad = 1 - \alpha + o(1).
\end{aligned}
$$

Thus, one potentially fruitful direction for inference for the value function is to develop smooth approximations to the value with a known or consistently estimable bound on the approximation error.

However, the truncated value proposed by GSZK need not satisfy (P1). We demonstrate this using a one-stage decision problem. The observed data are $\{(X_i, A_i, Y_i)\}_{i=1}^n$ where: $X \in \mathbb{R}^p$ denotes pre-treatment patient information; $A \in \{-1, 1\}$ denotes the treatment received; and $Y \in \mathbb{R}$ is the outcome coded so that higher values are better. Let $X_0 \in \mathbb{R}$ and $X_1 \in \mathbb{R}^p$ be known features of X and suppose that $Y = X_0\beta_0^* + AX_1^\intercal\beta_1^*/2 + \delta$ where: $\delta \sim \text{Normal}(0, \tau^2)$; and $X_0 = L1_{X_1^\intercal\beta^*=\epsilon} + M1_{X_1^\intercal\beta^*=0} + \eta$ for constants $M, L \in \mathbb{R}$ and $\eta \sim \text{Normal}(0, \gamma)$. Suppose that the postulated Q-function is $Q(x, a; \beta) = x_0\beta_0 + ax_1^\intercal\beta_1$ and thus is correctly specified. Let $\widehat{\beta} = \arg\min_\beta \mathbb{P}_n\{Y - Q(X, A; \beta)\}^2$. The estimated optimal policy is $\widehat{\pi}_n(x) = \text{sgn}(x_1^\intercal\widehat{\beta}_1)$; under standard regularity conditions $\sqrt{n}(\widehat{\beta} - \beta^*)$ is asymptotically normal. Assume A is randomly assigned with $P(A = 1|X) = P(A = -1|X) = 1/2$ with probability one, then under sufficient regularity conditions (Qian and Murphy, 2011; Zhao et al., 2012; Zhang et al., 2012) the value is $V(\widehat{\pi}_n) = 2PY1_{AX_1^\intercal\beta_1>0}$. For fixed $\epsilon > 0$ the truncated value is $V_\epsilon(\widehat{\pi}_n) = 2PY1_{AX_1^\intercal\widehat{\beta}_1>0}1_{|X_1^\intercal\widehat{\beta}_1|>\epsilon}$. Define $\widehat{V}_{\epsilon,n}(\widehat{\pi}_n) = 2\mathbb{P}_n1_{AX_1^\intercal\widehat{\beta}_1>0}1_{|X_1^\intercal\widehat{\beta}_1|>\epsilon}$, then

$$
\begin{aligned}
\sqrt{n}(\widehat{V}_{\epsilon,n}(\widehat{\pi}_n) - V_\epsilon(\widehat{\pi}_n)) &= 2\sqrt{n}(\mathbb{P}_n - P)Y1_{AX_1^\intercal\widehat{\beta}_1>0}1_{|X_1^\intercal\widehat{\beta}_n|>\epsilon}\\
&= 2\sqrt{n}(\mathbb{P}_n - P)Y1_{AX_1^\intercal\widehat{\beta}_1>\epsilon}\\
&= 2\sqrt{n}(\mathbb{P}_n - P)1_{AX_1^\intercal\sqrt{n}(\widehat{\beta}_1-\beta_1^*)>0}1_{AX_1^\intercal\beta_1^*=\epsilon} + o_p(1)
\end{aligned}
$$

where the last term is not asymptotically normal and is highly sensitive to the distribution of X_1 and value of β_1^*. Note also that if $X_1^\intercal\beta_1^* = 0$ with probability one then $V_\epsilon(\widehat{\pi}_n) - V(\widehat{\pi}_n) = V(\widehat{\pi}_n) = 2PM1_{AX_1^\intercal\widehat{\beta}_1>0} = M$ which can be made arbitrarily large (small); this underscores the need for a data-dependent bound on the approximation error.

Inspired by GSZK, we consider the alternative estimand $V_\Psi(\widehat{\pi}_n) = 2PY\Psi(AX_1^\intercal\widehat{\beta}_1)$ where $\Psi(u)$ is a continuously differentiable function used to approximate $1_{u>0}$. Furthermore, assume that the class of functions $\mathcal{F} = \{f(y, x, a; \beta_1) = y(\Psi(ax^\intercal\beta_1) - 1_{ax^\intercal\beta_1>0}) \,:\, \beta_1 \in \mathbb{R}^p\}$ is Glivenko-Cantelli (Kosorok, 2008).

A natural choice for Ψ is a sigmoid function though other choices are possible. Let $\widehat{V}_{\Psi}(\widehat{\pi}_n) = 2\mathbb{P}_n Y \Psi(AX_1^{\intercal}\widehat{\beta}_1)$; under mild moment assumptions a Taylor series argument shows that $\sqrt{n}(\widehat{V}_{\Psi}(\widehat{\pi}_n) - V_{\Psi}(\widehat{\pi}_n))$ is regular and asymptotically normal. The approximation error is bounded by $b(\epsilon) = |V_{\Psi}(\widehat{\pi}_n) - V(\widehat{\pi}_n)| + \epsilon$ which is consistently estimated by $\widehat{b}(\epsilon) = |\widehat{V}_{\Psi}(\widehat{\pi}_n) - \widehat{V}(\widehat{\pi})| + \epsilon = \big|\mathbb{P}_n Y(\Psi(AX_1^{\intercal}\widehat{\beta}_1) - 1_{AX_1^{\intercal}\widehat{\beta}_1 > 0})\big| + \epsilon$. We conjecture that the performance of this method will be sensitive to the choice of Ψ. If Ψ closely approximates the step function $1_{u>0}$ then the approximation error bound $b(\epsilon)$ will be small but it may be difficult to form a high-quality estimator of the sampling distribution of $\sqrt{n}(\widehat{V}_{\Psi}(\widehat{\pi}_n) - V_{\Psi}(\widehat{\pi}_n))$ because the derivative of Ψ must be very large near the origin; on the other hand, if the derivative of Ψ is not large near the origin then it may be possible to obtain a high-quality estimator of the sampling distribution of $\sqrt{n}(\widehat{V}_{\Psi}(\widehat{\pi}_n) - V_{\Psi}(\widehat{\pi}_n))$ but the approximation error bound may be large.

2. New directions for estimation and inference

2.1. Coherent confidence intervals

RR illustrate the perils of using separate confidence intervals for the decision boundary at each patient history rather than a single joint confidence interval for the entire decision boundary across all patient histories. In the RR example, there are three possible patient histories coded $x \in \{-1, 0, 1\}$, two treatments $a \in \{-1, 1\}$. Thus, there are eight regimes, each represented as a triple (a_{-1}, a_0, a_1), where $a_x \in \{-1, 1\}$ denotes the treatment assigned to a subject with $X = x$. In the RR example, the postulated model class for the decision boundary excludes the regime $(1, -1, 1)$ yet this regime could be implemented by a clinician who selects treatments using clinical judgment whenever a univariate confidence interval for the decision boundary at a given patient history contains zero. Thus, using separate confidence intervals for each patient history can lead to "incoherent" clinician behavior. However, instead of evaluating the performance of these confidence intervals by considering a population of clinicians, we argue that we should evaluate the performance across a population of patients. Thus, the question from our point of view is whether any patients are being treated inappropriately. A patient with history $x = -1$ is unaffected by what their treatment would have been had their history been 0 or 1. In the RR example suppose that *all* clinicians apply the "incoherent" regime $(1, -1, 1)$. Patients with $x = 1$ receive the (estimated) optimal treatment $a = 1$; patients with $x = -1$ receive $a = 1$ which is not dominated by $a = -1$ and is thus appropriate; and similarly, patients with $x = 0$ receive $a = -1$ which is not dominated by $a = 1$ and thus appropriate.

However moving to the multi-stage setting we appreciate RR's point of view in regards to potentially incoherent sequences of treatments being assigned to a patient. Indeed some of us, Laber et al. (2014) define the set of *feasible* regimes as those that are consistent with a partial ordering induced on the treatments (e.g., a partial ordering induced by separate confidence intervals for each patient history) and representable in the postulated model space (they did not explore incoherence further). We and others are currently working on this problem.

2.2. Soft-max Q-learning

GSZK propose a soft-max Q-learning as an alternative to Q-learning. Because soft-max Q-learning involves only smooth operations of the data standard methods for inference apply. We show that soft-max Q-learning can be viewed as regular Q-learning applied to a stochastic policy in which the propensity of the clinician to follow the estimated optimal second stage policy varies with the estimated second stage effect size. Soft-max Q-learning uses predicted outcome $\check{Y} = H_{2,0}^{\intercal}\widehat{\beta}_{2,0} + \alpha^{-1}\log(1 + \exp\{\alpha H_{2,1}^{\intercal}\widehat{\beta}_{2,1}\})$. It can be shown that $\sup_{v\in\mathbb{R}} |\alpha^{-1}\log(1+\exp\{\alpha v\}) - \text{expit}(\alpha v)v| = o(1/\alpha)$ as $\alpha \to \infty$, where $\text{expit}(u) = \exp(u)/(1+\exp(u))$. Let $\mathcal{D}$ denote the observed data and let $\widehat{\pi}_2^{\alpha}(h_2)$ be a stochastic second stage policy that satisfies $P(\widehat{\pi}_2^{\alpha}(h_2) = \widehat{\pi}_2(h_2)|H_2 = h_2, \mathcal{D}) = \text{expit}(\alpha H_{2,1}^{\intercal}\widehat{\beta}_{2,1})$. A clinician acting according to $\widehat{\pi}_2^{\alpha}$ is increasingly more likely to recommend treatments consistently with the Q-learning estimated optimal second stage decision rule as the magnitude of the estimated second stage effect size, $H_{2,1}^{\intercal}\widehat{\beta}_{2,1}$, increases. Then, $\mathbb{E}(\widehat{Q}_2(H_2, \widehat{\pi}_2^{\alpha}(H_2))|H_2, \mathcal{D}) = H_{2,0}^{\intercal}\widehat{\beta}_{2,0} + \text{expit}(H_{2,1}^{\intercal}\widehat{\beta}_{2,1})H_{2,1}^{\intercal}\widehat{\beta}_{2,1} \approx \check{Y}$. Thus, soft-max Q-learning can be viewed as estimating the optimal first stage decision rule assuming the clinician will follow a stochastic policy described by $\widehat{\pi}_2^{\alpha}$ at the second stage. Because soft-max Q-learning is smooth, it should possible to conduct the conditional inference for the first stage Q-function as suggested by RR; such conditional inference would therefore accommodate not only the data-driven decision rule at the second stage (as suggested by RR) but also uncertain clinician behavior.

2.3. Variable screening for SMARTs

Hsu and Small (HS hereafter) propose a method for identifying variables that may be of interest for follow-up investigation within the context of a SMART. In the setup considered by HS the observed data on each subject are (A, D, Y) where: $A \in \{0,1\}$ is a randomized treatment; $D \in \{0,1\}$ is an intermediate (post-treatment) outcome; and $Y \in \mathbb{R}$ is a distal outcome. For simplicity, assume that Y is binary and coded to take values in $\{0,1\}$. HS's idea is to use D as an indicator that a subject's initial treatment should have been switched; thus, in a future study upon observing D the clinician has the opportunity to switch from treatment a to treatment $1-a$. Let $\{(Y^{(a=j)}, D^{(a=j)})\}_{j=0,1}$ denote the set of potential outcomes. Define $C_j^k = P(Y^{(a=k)} = 1, Y^{(a=1-k)} = 0|D^{(a=k)} = j) - P(Y^{(a=k)} = 0, Y^{(a=1-k)} = 1|D^{(a=k)} = j)$ for $j, k \in \{0,1\}$. In this setting (1) in HS is equivalent to $C_1^1 > C_0^1$. For clarity we describe this in words. (1) only considers a population of individuals who start off on treatment, $a = 1$. $C_1^1 > 0$ means that among the subpopulation of individuals experiencing $D = 1$ in response to $a = 1$, a higher fraction have started off on their optimal treatment than have started off on their non-optimal treatment. Similarly $C_0^1 > 0$ means that among the subpopulation experiencing $D = 0$ in response to $a = 1$, a higher fraction have started off on their non-optimal treatment than have started off on their optimal treatment. Thus $C_1^1 > C_0^1$ is a comparison of differences in fractions, one difference per subpopulation. Among those who experience $D = 1$ the difference in fractions of individuals who started off on their optimal treatment versus did not start off on their optimal treatment is higher than the same fraction among those who experience $D = 0$.

At first it was unclear to us how (1) in HS's discussion might be used to inform treatment decisions. However suppose we are willing to assume no carry-over effect. That is, if a patient switches treatment then the patient's long run outcome Y will only depend on the new treatment. In this case consider instead testing if $C_j^1 C_j^0 < 0$ for each $j = 0, 1$; if the test indicates that $C_j^1 C_j^0 < 0$, assuming that there is no carry-over effect, D could be used with future patients to dictate a treatment switch. If the estimated $\hat{C}_j^k < 0$ but $\hat{C}_j^{1-k} > 0$ then future patients initially treated with $a = k$ experiencing intermediate outcome $D = j$ might be switched to $a = 1 - k$. Thus, in this sense, the method proposed by HS can be used to inform treatment decisions.

3. Scalability and the definition of a direct estimator

3.1. Scaling estimation an inference to large problems

Banerjee (B hereafter) asked about computation and asked how the methods and technical issues scale to settings with high-dimensional or continuous actions. Computation of the bounds requires solving a non-convex optimization problem of the form

$$\min_{\gamma \in \mathbb{R}^p} \sum_{i=1}^{n} \omega_i \left([r_i^{\intercal}\gamma + d_i]_+ - [r_i^{\intercal}\gamma + e_i]_+ \right), \tag{1}$$

where $\omega_1, \ldots, \omega_n, d_1, \ldots, d_n, e_1, \ldots, e_n \in \mathbb{R}$ and $r_1, \ldots, r_n \in \mathbb{R}^p$ are known fixed constants. In our simulations the dimension of γ was sufficiently small so that we could compute an approximate solution using a stochastic search; the search was tuned to ensure a sufficient number of points were evaluated before termination. However, there are several other approaches that could be used to compute an approximate solution to (1). One approach that scales well to high-dimensional (large p) problems is coordinate descent. Note that (1) is continuous and piecewise linear. Suppose first that $p = 1$. In this case it can be seen that any optimal solution γ^* must solve $r_i\gamma^* + d_i = 0$ or $r_i\gamma^* + e_i = 0$ for some $1 \leq i \leq n$. Thus, there are at most $2n$ candidate solutions that must be examined to find an optimal solution in the $1d$ case. For any $v \in \mathbb{R}^p$ let $v_{(j)} = (v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_p)$ and define

$$\widehat{\gamma}_j(v_{(j)}) = \arg\min_{\tau \in \mathbb{R}} \sum_{i=1}^{n} \omega_i \left(\Big[r_{ij}\tau + r_{i(j)}^{\intercal} v_{(j)} + d_i \Big]_+ - \Big[r_{ij}\tau + r_{i(j)}^{\intercal} v_{(j)} + e_i \Big]_+ \right). \tag{2}$$

Computing (2) is equivalent to solving (1) with $p = 1$ and thus requires evaluation of $2n$ potential solutions. The coordinate descent algorithm is as follows: (i) initialize γ to a starting value in $\mathbb{R}^p$; (ii) repeat updates of the form $\gamma_j = \widehat{\gamma}_j(\gamma_{(j)})$ cycling over $j = 1, \dots, p$ until changes in γ are sufficiently small. Coordinate descent is easily to implement and the evaluation of the $2n$ potential solutions required for each update is trivial to parallelize; we have had success with this approach in other contexts with similar non-differentiable but continuous objectives (Laber and Murphy, 2013).

An alternative approach to computing (1) is to recognize that it can be written as

$$\inf_{\gamma \in \mathbb{R}^p} \sum_{i=1}^{n} \Big\{ \big([\omega_i]_+ \, [r_i^\intercal \gamma + d_i]_+ + [\omega_i]_- \, [r_i^\intercal \gamma + e_i]_+ \big) \\ - \big([\omega_i]_- \, [r_i^\intercal \gamma + d_i]_+ + [\omega_i]_+ \, [r_i^\intercal \gamma + e_i]_+ \big) \Big\}$$

which is the difference of two convex functions and thus (1) is a DC programming problem. Hence, one can apply DC algorithms (see Horst and Thoai, 1999, and references therein) to approximate (1). The DC algorithm has a track record of being effective in large problems but our experience with this algorithm is limited.

B also asked about nonregularity in the contexts of continuous treatments. This could arise in the context of clinical practice, say when optimizing dose. To illustrate how nonregularity can occur in the context of a continuous treatments suppose that $a_2 \in [-d, d]$ and consider a nonlinear model for the second stage Q-function of the form

$$Q_2(h_2, a_2; \beta_2, \sigma) = h_{2,0}^\intercal \beta_{2,0} + \exp\left\{ (a_2 - h_{2,1}^\intercal \beta_{2,1})^2 / \sigma + h_{2,2}^\intercal \beta_{2,2} \right\}, \qquad (3)$$

where $h_{2,0}, h_{2,1}, h_{2,2}$ are known features of h_2. In (3) the optimal dosage is the projection of $h_{2,1}^\intercal \beta_{2,1}$ onto $[-d, d]$, and σ, $h_{2,2}^\intercal \beta_{2,2}$ govern the treatment effect size. Let $\widehat{\beta}_2, \widehat{\sigma}$ denote the non-linear least squares estimators of the parameters indexing the working second-stage Q-function. The maximized estimated second stage Q-function is

$$\sup_{a_2 \in [-d,d]} Q_2(h_2, a_2; \widehat{\beta}_2, \widehat{\sigma}) = h_{2,0}^\intercal \widehat{\beta}_{2,0} + \exp\left\{ h_{2,2}^\intercal \widehat{\beta}_{2,2} \right\} 1_{|h_{2,1}^\intercal \widehat{\beta}_{2,1}| \le d} \\ + \exp\left\{ (d - |h_{2,1}^\intercal \widehat{\beta}_{2,1}|)^2 / \widehat{\sigma} + h_{2,2}^\intercal \widehat{\beta}_{2,2} \right\} 1_{|h_{2,1}^\intercal \widehat{\beta}_{2,1}| > d},$$

which is not smooth in the data. Thus, non-regularity can occur even with continuous treatments.

A related problem raised by B is a discrete but high-dimensional treatment. However, unlike the continuous treatment setting, with a high-dimensional discrete treatment there may be no notion of smoothness across values of treatment thereby making it difficult to pool information across units of observation. Without massive amounts of data additional structure must be introduced into the working models. One setting in which this occurs is when treatments are made at each time point across a series of spatial locations. Suppose for simplicity that one can apply one of two treatments, say 0 or 1 at each of N locations; in this setting $a \in \{0, 1\}^N$. However, it may be reasonable to assume that a treatment applied at a given location diffuses quickly over space and that any heterogeneity in the treatment effect over space can be modeled parametrically using observable covariates at each location.

3.2. Definition of a direct estimator

RR correctly note that the estimator that we discuss in Appendix A should not be viewed as providing an estimator of the optimal DTR. Note that our asymptotic results describe the behavior of the parameter estimators about their limiting values and do not refer to an optimal DTR. Unfortunately, this was not clear in our description.

We believe that in some settings there are other, equally valid or better, criteria for assessing the quality of direct estimators than consistent estimation of the parameters indexing the estimated optimal DTR. In some settings, constructing a high-quality DTR from within a pre-specified class, e.g., restricted to be parsimonious, logistically feasible, low-cost, etc., may be more of a priority than constructing a consistent estimator of π^{opt} (Orellana et al., 2010; Zhang et al., 2012, 2013). In such cases, performance guarantees in the form of error bounds, i.e., bounds on the difference between the value of an estimated policy and the value of π^{opt}, may be more appropriate. Error bounds have been used extensively in computer science (e.g., Bartlett et al., 2006, and references therein) where the focus is primarily on performance rather than making statements about the true structure of the optimal DTR. For example, Zhao et al. (2012) derive error bounds on a direct estimator of the form used in Appendix A.

References

Bartlett, P.L., Jordan, M.I., and McAuliffe, J.D., Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. MR2268032

Horst, R. and Thoai, N.V., Dc programming: Overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, 1999. MR1715016

Kosorok, M.R., *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008. MR2724368

Laber, E.B. and Murphy, S.A., Adative inference after model selection, 2013.

Laber, E.B., Lizotte, D.J., and Ferguson, B. Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*, 2014.

Orellana, L., Rotnitzky, A., and Robins, J., Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: Main content. *Int. Jrn. of Biostatistics*, 6(2), 2010. MR2602551

QIAN, M. and MURPHY, S.A., Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39(2):1180–1210, 2011. MR2816351

SCHULTE, P.J., TSIATIS, A.A. LABER, E.B., and DAVIDIAN, M., Q- and a-learning methods for estimating optimal dynamic treatment regimes. Technical Report, 1202.4177v2, arXiv.org, 2013.

ZHANG, B., TSIATIS, A.A., LABER, E.B., and DAVIDIAN, M., Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, To appear, 2013. MR3094445

ZHANG, B., TSIATIS, A.A., LABER, E.B., and DAVIDIAN, M., A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012. MR3040007

ZHAO, Y., ZENG, D., RUSH, A.J., and KOSOROK, M.R., Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012. MR3010898