Interactive Q-Learning

Eric B. Laber

Department of Statistics, North Carolina State University

ENAR, March 11, 2013

Acknowledgments

Thanks to

- Min Zhang
- NCSU and UNC DTR Working Groups
- All of you!

Joint work with

- Len Stefanski
- Kristin Linn

Dynamic treatment regimes

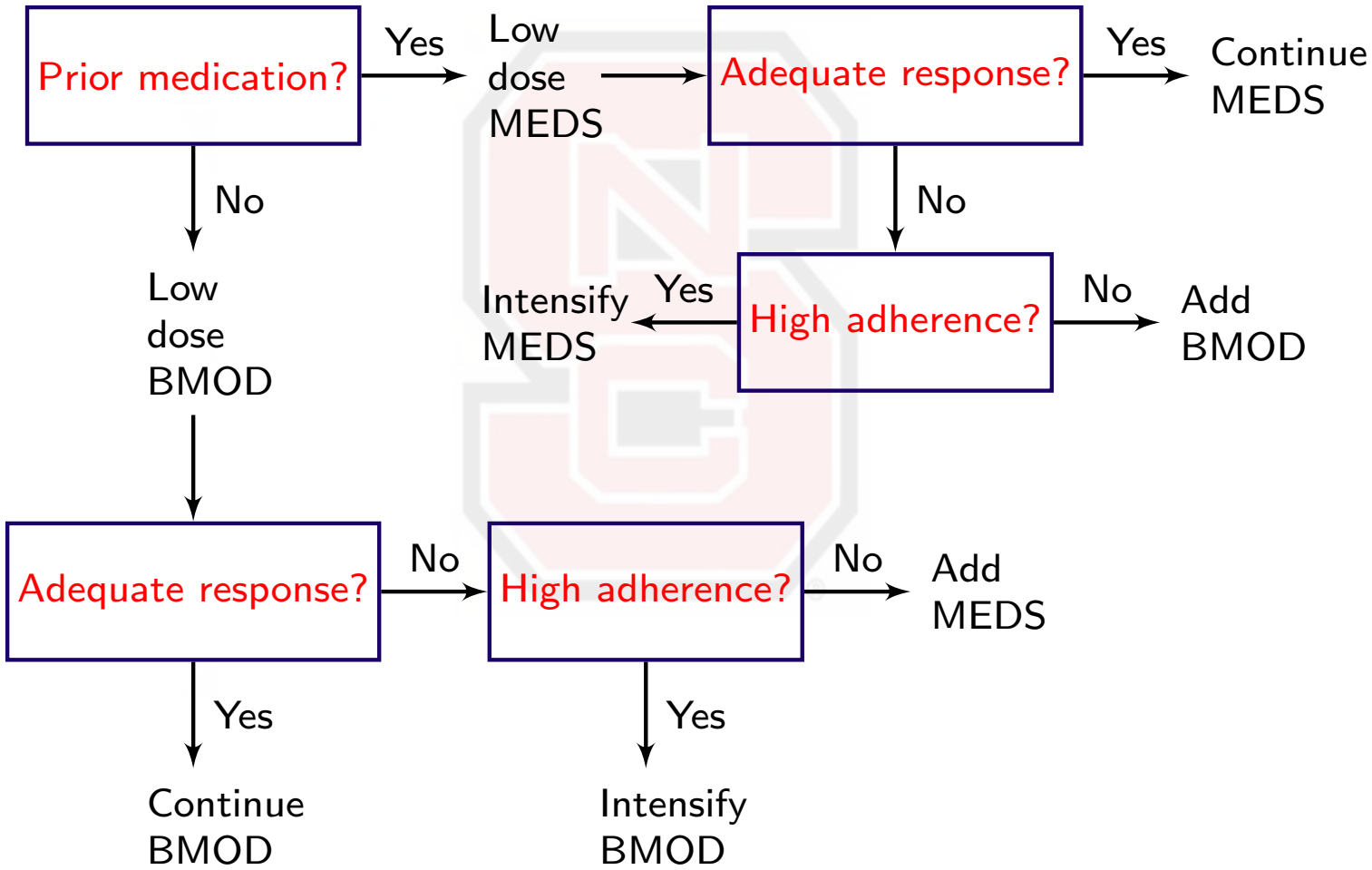
- Motivation : treatment of chronic illness
  - Some examples: HIV/AIDS, cancer, depression, schizophrenia, drug and alcohol addiction, ADHD, etc.
  - Multistage decision making problem
  - Longer-term treatment requires consideration and tradeoff of present versus longer term benefit.

- Dynamic treatment regimes (DTRs)
  - Operationalize multistage decision making via as sequence of decision rules
    - One decision rule for each stage of intervention
    - A decision rule maps up-to-date patient information to a recommended treatment
  - Aim to optimize some cumulative clinical outcome

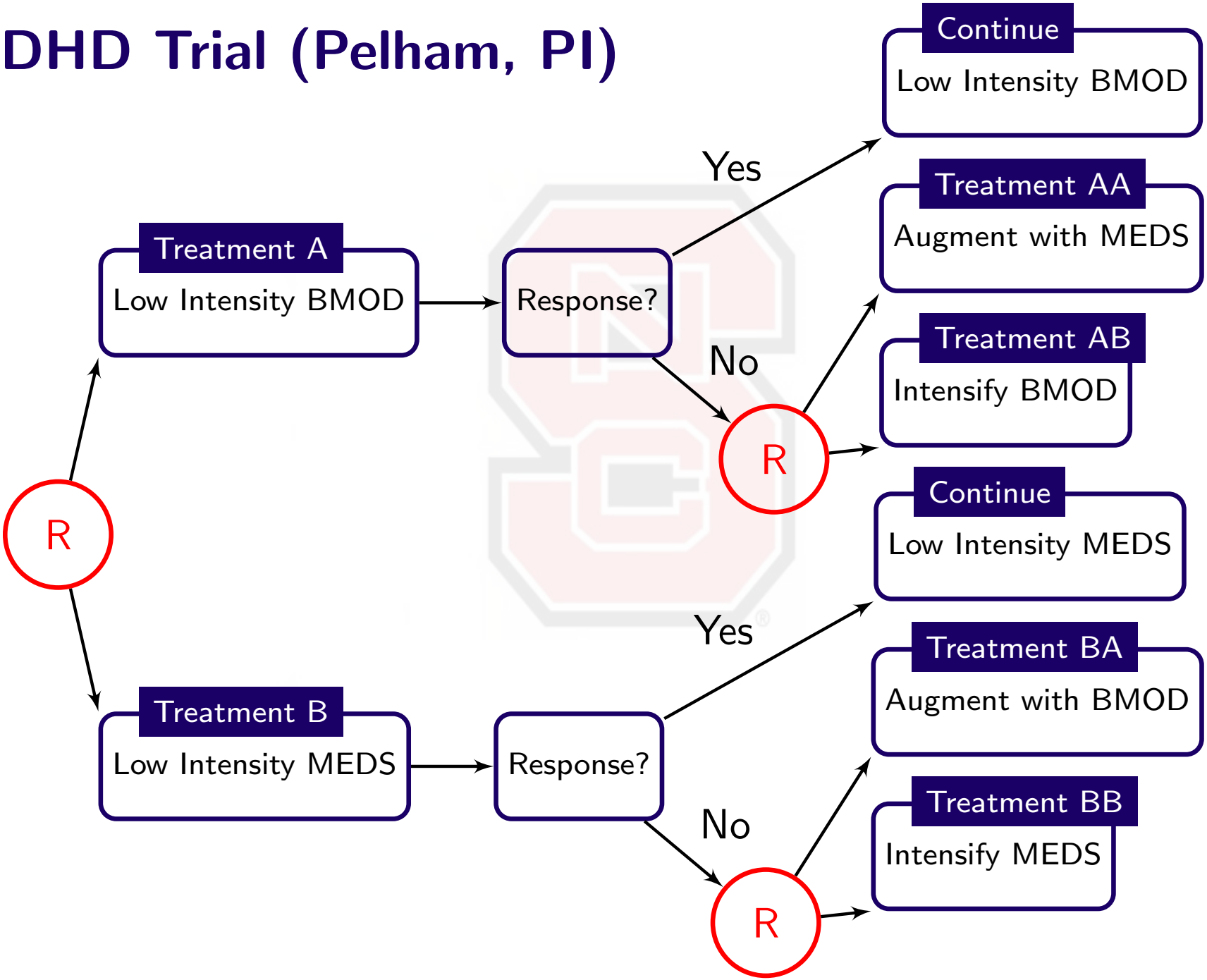
Other applications

- Construction and inference for policies have applications beyond medicine
  1. Artificial Intelligence and Reinforcement Learning (autonomous helicopter, drones, etc., Ng 2003)
  2. Marketing (Simester, Sun and Tsitsiklis, 2003)
  3. Active labor market policies (Lechner and Miquel, 2010)
  4. ...

An example DTR for ADHD



ADHD Trial (Pelham, PI)



Data and DTRs

- Restrict attention to two-stage randomized trials with binary treatments at each stage
- (X_1, A_1, X_2, A_2, Y) for each individual
  - X_j: Stage j patient information
  - A_j: Treatment at stage j
  - Y: Primary outcome (larger is better)
  - H_j: History at stage j, $H_1 = X_1,\ H_2 = (X_1, A_1, X_2)$
- The regime, $\pi = \{\pi_1, \pi_2\},\ \pi_j : \mathrm{dom}(H_j) \rightarrow \mathrm{dom}(A_j)$, should have high Value: $V^\pi = E^\pi(Y)$
  - The value corresponds to the average outcome if all patients are assigned treatment according to π

Review of dynamic programming

- If generative model known, optimal policy follows from dynamic programming $\pi^{dp} = (\pi_1^{dp}, \pi_2^{dp})$
  1. Define $Q_2(h_2, a_2) \triangleq \mathbb{E}(Y|H_2 = h_2, A_2 = a_2)$ and subsequently $\pi_2^{dp}(h_2) = \arg\max_{a_2} Q_2(h_2, a_2)$
  2. Define expected outcome following optimal second stage decision rule $Y^* \triangleq \max_{a_2} Q_2(H_2, a_2)$
  3. Define $Q_1(h_1, a_1) \triangleq \mathbb{E}(Y^*|H_1 = h_1, A_1 = a_1)$ and subsequently $\pi_1^{dp}(h_1) = \arg\max_{a_1} Q_1(h_1, a_1)$

Approximate dynamic programming: Q-learning

- Underlying generative distribution is unknown
- Approximate conditional expectations with regressions
- Linear regressions; $A_j \in \{-1, 1\}$, H_{j1}, H_{j2} features of patient history H_j
  1. Regress Y on H_{21}, H_{22} to obtain $\hat{Q}_2(H_2, A_2) = \hat{\beta}_{21}^\intercal H_{21} + \hat{\beta}_{22}^\intercal H_{22} A_2$ and subsequently $\hat{\pi}_2(h_2) = \arg\max_{a_2} \hat{Q}_2(h_2, a_2)$
  2. Define the estimated outcome following the optimal second stage decision rule $\hat{Y} = \max_{a_2} \hat{Q}_2(H_2, a_2)$
  3. Regress $\hat{Y}$ on H_{11}, H_{12} to obtain $\hat{Q}_1(H_1, A_1) = \hat{\beta}_{11}^\intercal H_{11} + \hat{\beta}_{12}^\intercal H_{12} A_1$ and subsequently $\hat{\pi}_1(h_1) = \arg\max_{a_1} \hat{Q}_1(h_1, a_1)$

Approximate dynamic programming: Q-learning

- Underlying generative distribution is unknown
- Approximate conditional expectations with regressions
- Linear regressions; $A_j \in \{-1, 1\}$, H_{j1}, H_{j2} features of patient history H_j

Modeling

1. Regress Y on H_{21}, H_{22} to obtain $\hat{Q}_2(H_2, A_2) = \hat{\beta}_{21}^\intercal H_{21} + \hat{\beta}_{22}^\intercal H_{22} A_2$ and subseq. $\hat{\pi}_2(h_2) = \arg\max_{a_2} \hat{Q}_2(h_2, a_2)$

Maximization

2. Define the estimated outcome following the optimal second stage decision rule $\hat{Y} = \max_{a_2} \hat{Q}_2(H_2, a_2)$

Modeling

3. Regress $\hat{Y}$ on H_{11}, H_{12} to obtain $\hat{Q}_1(H_1, A_1) = \hat{\beta}_{11}^\intercal H_{11} + \hat{\beta}_{12}^\intercal H_{12} A_1$ and subseq. $\hat{\pi}_1(h_1) = \arg\max_{a_1} \hat{Q}_1(h_1)$

Q-learning in practice

- Theoretical problems
  - Nonsmooth, nonmonotone max-operator leads to nonregular limit theory
  - Standard asymptotic approaches (e.g. bootstrap, normal approx, etc.) do not apply without modification
  - Inference is notoriously difficult
- Practical problems
  - Linear models are misspecified under simple and realistic generative models
  - Difficult to correctly capture the functional form of the transformed data even with flexible models
  - Model building and validation would require either developing specialized tools or ignoring the foregoing problems

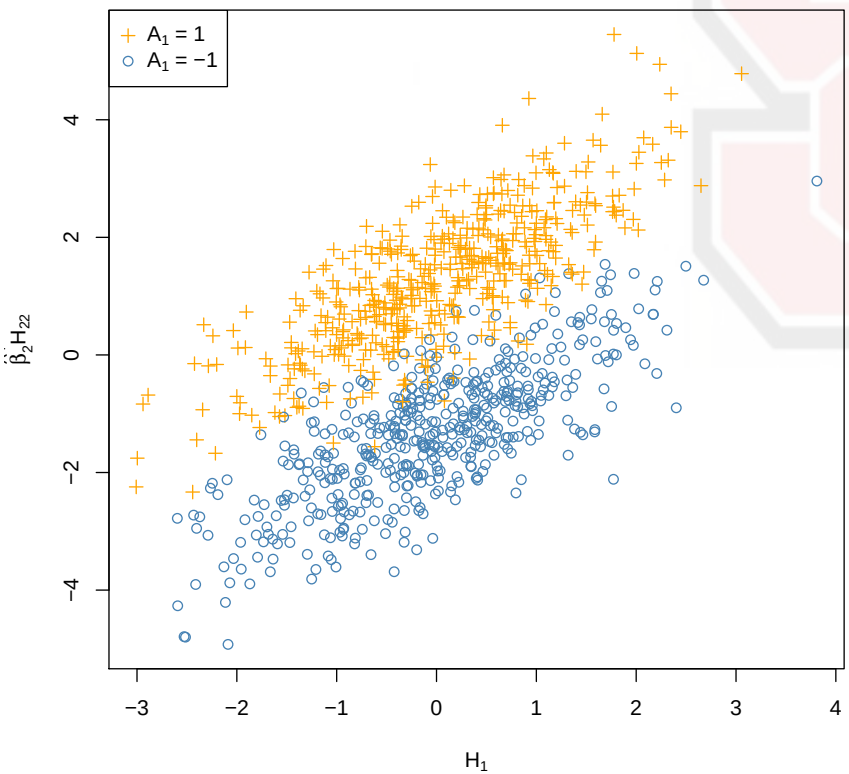
Q-learning in practice, cont'd

- Recall $\hat{Y} = \max_{a_2} \hat{Q}(H_2, a_2) = \hat{\beta}_{21}^{\mathsf{T}} H_{21} + |\hat{\beta}_{22}^{\mathsf{T}} H_{22}|$

Q-learning in practice, cont'd

- Recall $\hat{Y} = \max_{a_2} \hat{Q}(H_2, a_2) = \hat{\beta}_{21}^{\mathsf{T}} H_{21} + |\hat{\beta}_{22}^{\mathsf{T}} H_{22}|$

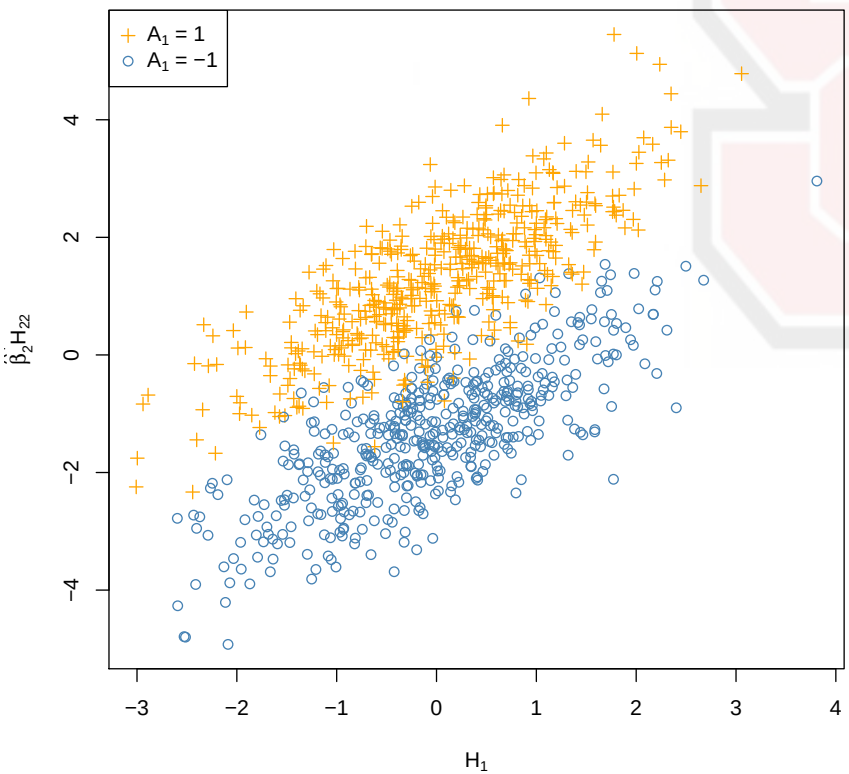
Before maximization



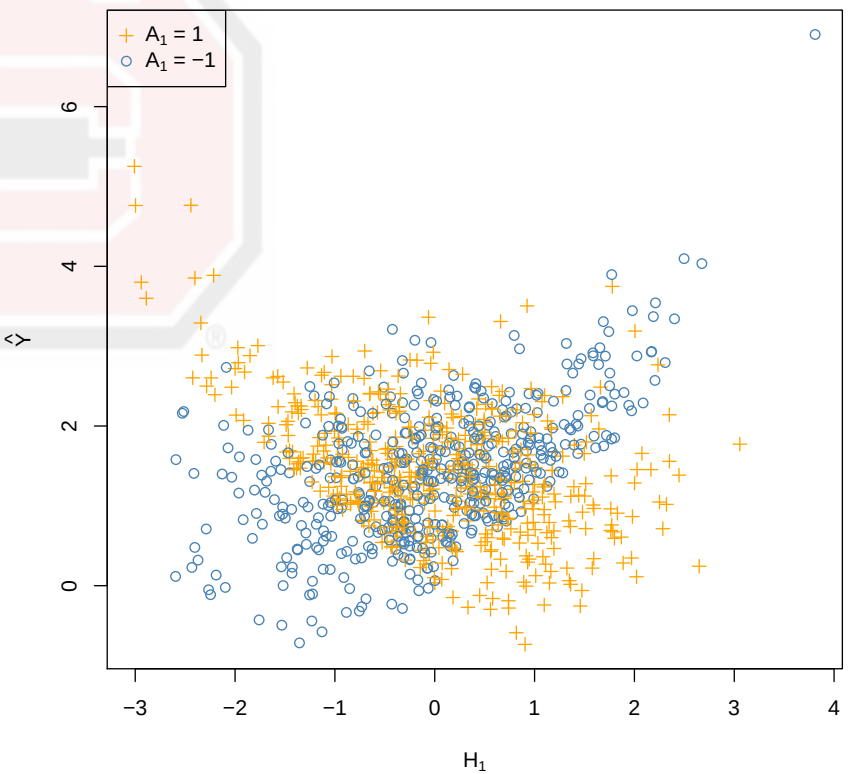
Q-learning in practice, cont'd

- Recall $\hat{Y} = \max_{a_2} \hat{Q}(H_2, a_2) = \hat{\beta}_{21}^{\mathsf{T}} H_{21} + |\hat{\beta}_{22}^{\mathsf{T}} H_{22}|$

Before maximization



After maximization



IQ-learning

- Is it possible to interchange the maximization and modeling step so that all modeling takes place *before* the maximization is applied?

IQ-learning

- Is it possible to interchange the maximization and modeling step so that all modeling takes place *before* the maximization is applied?
- Recall that $Q_1(H_1, A_1) = \mathbb{E}(\max_{a_2} Q_2(H_2, A_2)|H_1, A_1)$ which we can write as

$$\mathbb{E}\left(\beta_{21}^{*\intercal} H_{21} \middle| H_1, A_1\right) + \mathbb{E}\left(|\beta_{22}^{*\intercal} H_{22}| \middle| H_1, A_1\right)$$
$$= \mathbb{E}\left(\beta_{21}^{*\intercal} H_{21} \middle| H_1, A_1\right) + \int |z| f(z|H_1, A_1) dz,$$

where $f(z|H_1, A_1)$ denotes the conditional density of $\beta_{22}^{*\intercal} H_{22}|H_1, A_1$.

IQ-learning cont'd

- Idea! Model f and $\mathbb{E}(\beta_{21}^{*\mathsf{T}} H_{21} | H_1, A_1)$
  - Only involves smooth (linear) transformations of the data
  - Standard modeling building techniques apply
  - Simple normal limit theory
- No free lunch
  - Modeling a conditional density is more involved than modeling conditional expectations
  - However, mean-variance models may be sufficiently expressive and are easy to construct

Mean-variance models for f

- Normal linear location-scale estimator

$$f^N(z|h_1, a_1) = \frac{1}{\sigma}\phi\left(\frac{z - \theta_{11}^\intercal H_{11} - \theta_{12}^\intercal H_{12} A_1}{\sigma}\right)$$

- Many extensions
  - Flexible estimator for residual distribution (e.g., replace ϕ with non- or semi-parametric estimator)
  - More flexible mean and variance models

Generic IQ-learning algorithm

1. Regress Y on H_{21}, H_{22}, and A_2 to obtain $\hat{Q}_2(H_2, A_2) = \hat{\beta}_{21}^{\intercal} H_{21} + \hat{\beta}_{22}^{\intercal} H_{22} A_2$
2. Use $\{(\hat{\beta}_{22}^{\intercal} H_{22,i}, H_{1,i}, A_{1,i})\}_{i=1}^{n}$ to obtain an estimator $\hat{f}_n(.|H_1, A_1)$ of $f(.|H_1, A_1)$
3. Regress $\hat{\beta}_{21}^{\intercal} H_{21}$ on H_1, and A_1 to obtain an estimator $\hat{g}_n(H_1, A_1)$ of $\mathbb{E}(\beta_{21}^{*\intercal} H_{22}|H_1, A_1)$
4. Combine the above estimators to form

$$\hat{Q}_1^{IQ}(H_1, A_1) = \hat{g}_n(H_1, A_1) + \int |z| \hat{f}_n(z|H_1, A_1) dz.$$

Generic IQ-learning algorithm

Modeling 1. Regress Y on H_{21}, H_{22}, and A_2 to obtain $\hat{Q}_2(H_2, A_2) = \hat{\beta}_{21}^{\intercal} H_{21} + \hat{\beta}_{22}^{\intercal} H_{22} A_2$

Modeling 2. Use $\{(\hat{\beta}_{22}^{\intercal} H_{22,i}, H_{1,i}, A_{1,i})\}_{i=1}^{n}$ to obtain an estimator $\hat{f}_n(.|H_1, A_1)$ of $f(.|H_1, A_1)$

Modeling 3. Regress $\hat{\beta}_{21}^{\intercal} H_{21}$ on H_1, and A_1 to obtain an estimator $\hat{g}_n(H_1, A_1)$ of $\mathbb{E}(\beta_{21}^{*\intercal} H_{22}|H_1, A_1)$

Maximization 4. Combine the above estimators to form

$$\hat{Q}_1^{IQ}(H_1, A_1) = \hat{g}_n(H_1, A_1) + \int |z| \hat{f}_n(z|H_1, A_1) dz.$$

Generic IQ-learning algorithm

Modeling	1. Regress Y on H_{21}, H_{22}, and A_2 to obtain $\hat{Q}_2(H_2, A_2) = \hat{\beta}_{21}^{\mathsf{T}} H_{21} + \hat{\beta}_{22}^{\mathsf{T}} H_{22} A_2$
Modeling	2. Use $\{(\hat{\beta}_{22}^{\mathsf{T}} H_{22,i}, H_{1,i}, A_{1,i})\}_{i=1}^{n}$ to obtain an estimator $\hat{f}_n(.\|H_1, A_1)$ of $f(.\|H_1, A_1)$
Modeling	3. Regress $\hat{\beta}_{21}^{\mathsf{T}} H_{21}$ on H_1, and A_1 to obtain an estimator $\hat{g}_n(H_1, A_1)$ of $\mathbb{E}(\beta_{21}^{*\mathsf{T}} H_{22}\|H_1, A_1)$
Maximization	4. Combine the above estimators to form

$$\hat{Q}_1^{IQ}(H_1, A_1) = \hat{g}_n(H_1, A_1) + \int |z| \hat{f}_n(z|H_1, A_1) dz.$$

Modeling now involves only smooth functionals of the data!

Small simulation study

- We consider the following class of generative models

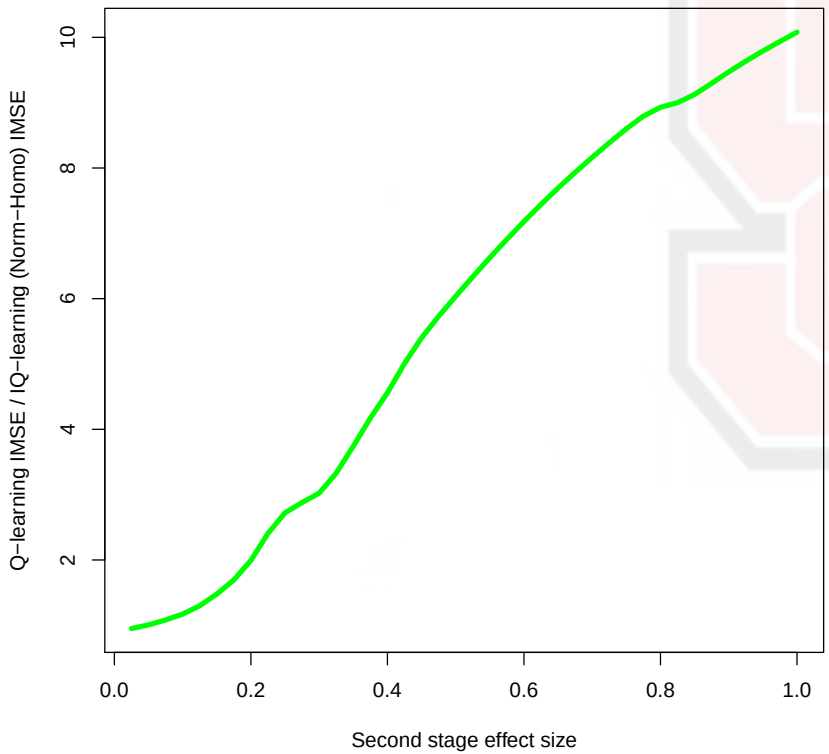
$$
\begin{aligned}
&X_1 \sim \text{Normal}_p(\mu 1_p, \Omega_{\text{AR}}(r)), \quad \xi \sim \text{Normal}_p(0, I_p) \\
&A_t \sim \text{Uniform}\{-1, 1\}, \qquad\qquad X_2 = ((A_1/2 + 3/2)X_1 + \xi, \\
&\phi \sim \text{Normal}(0, \gamma^2), \qquad\qquad Y = \beta_{21}^{\mathsf{T}} H_{21} + \beta_{22}^{\mathsf{T}} H_{22} A_2 + \phi.
\end{aligned}
$$

- Mimic SMART study with random treatment assignment
- Second stage Q-function is correctly specified
- We will fix all parameters except that we will vary the magnitude of β_{21} and β_{21}
- We use linear working models and assume normal residuals for IQ-learning estimator

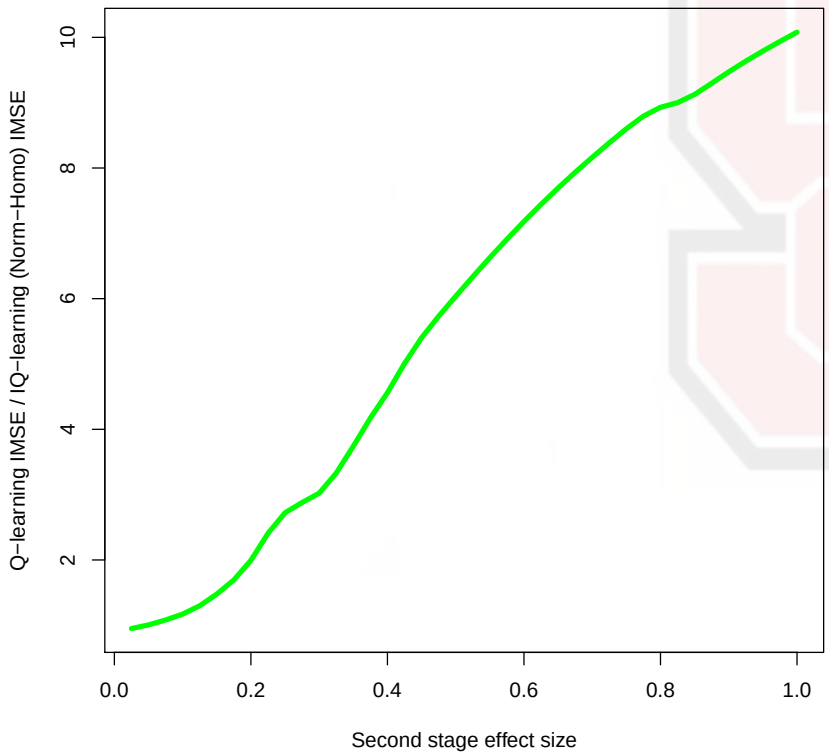
Small simulation study cont'd

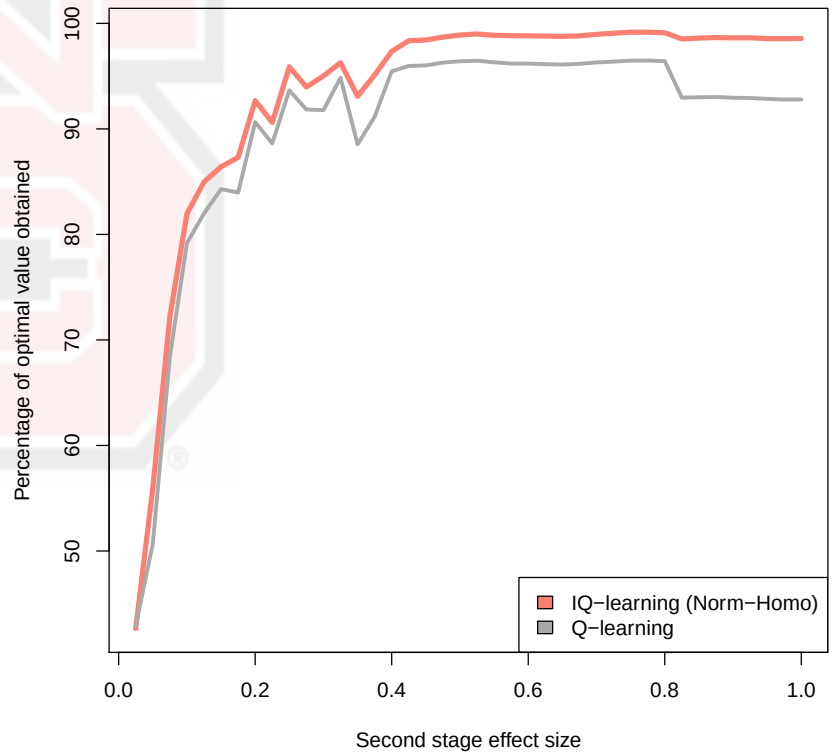
- Parameter settings
  - $r = .5$, $\mu = .1$, $\gamma^2 = 1$,
  - $\beta_{2,0} = \beta_{2,1} \propto (-1^{\intercal}_{p/2}, 4\,1^{\intercal}_{p/2})^{\intercal}$
  - Vary the second stage effect size by scaling $\beta_{2,0}$ and $\beta_{2,1}$
  - Training set size $n = 250$
  - 100 Monte Carlo replications
- Compare performance of *IQ*-learning and *Q*-learning
  - IMSE for estimating $Q_1(h_1, a_1)$
  - Value obtained (e.g., expected outcome if patients are assigned treatments using learned policy)
  - Estimated coverage of bootstrap confidence interval for $Q_1(h_1, a_1)$

Small simulation study: IMSE and value obtained

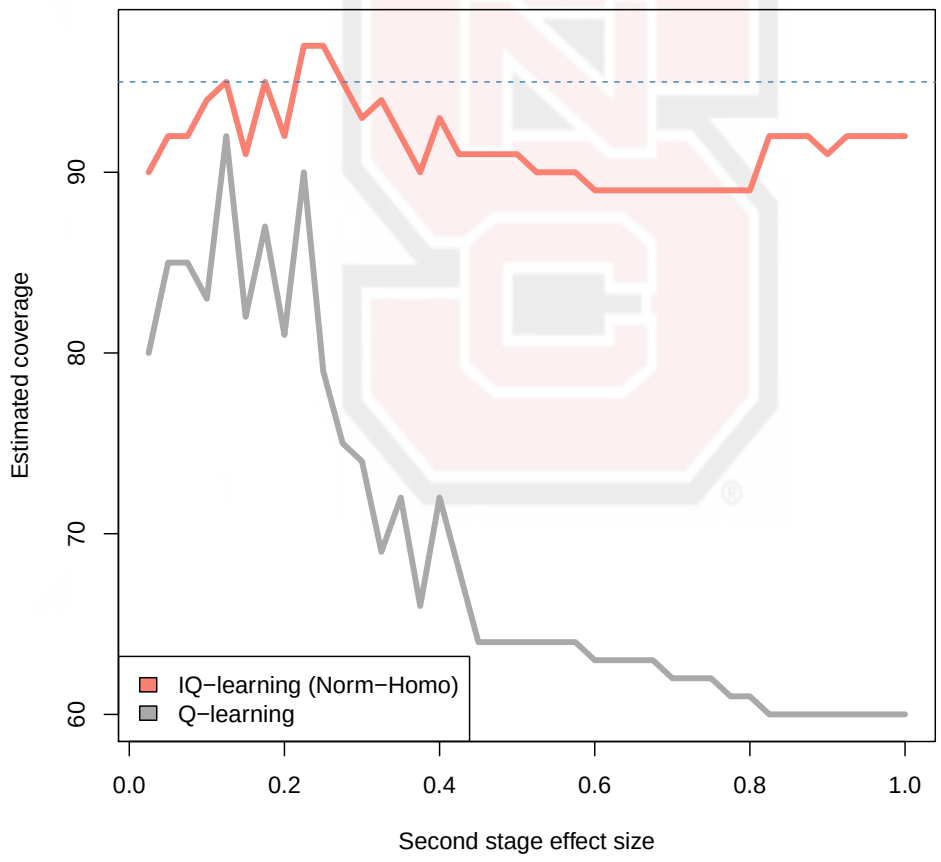


Small simulation study: IMSE and value obtained





Small simulation study: coverage



Small simulation study: conclusions

- Large improvement in IMSE for first stage Q-function
- General improvement for value, practically significant for larger effect sizes
- Coverage improves and is not signifantly below nominal levels for the range of effect sizes

Conclusions

- Summary
  - Proposed a novel method for estimating optimal DTRs that avoids modeling non-smooth transformations of the data
    - Mimics Q-learning but interchanges modeling and maximization
    - Correctly specified for a larger class of generative models than Q-learning
  - Simple asymptotic inference (not shown here)
  - Easier model building and critique
- Discussion
  - Did we solve the problem of nonregularity?
    - No. The first-stage Q-function is a nonsmooth functional of the underlying generative distribution. There are no regular or asymptotically unbiased estimators.
    - Our estimators are regular and asymptotically normal under mild conditions, but may not be consistent for some generative models.

Thank you for your attention

Questions: laber@stat.ncsu.edu
A copy of this talk can found at: www4.stat.ncsu.edu/∼laber

Acknowledgements:
National Institute of Health grant P50 DA10075