Missing Data and Uncertainty in Batch Reinforcement Learning

D. Lizotte, L. Gunter, E. Laber and S. A. Murphy
{danjl,lgunter,laber,samurphy}@umich.edu

University of Michigan
Department of Statistics
Institute for Social Research

The Problem

In the **batch, finite-horizon reinforcement learning (RL)** setting, **we estimate the optimal action value function**, i.e. the expected sum of future rewards one would obtain by taking an action having seen an observation and acting optimally thereafter.

In this setting, **we have a training set of sample trajectories**. The length of these trajectories is bounded by the horizon.

We are interested in the case where **some of the elements of the sample trajectories are missing**. This missingness could involve any of the observations, actions, or rewards in the training set, and could be different from trajectory to trajectory. Typical methods like linear regression cannot be directly applied.

We assume **the data are *not* Missing Completely At Random (MCAR.)** This means, for example, that **we cannot simply omit the incomplete sample trajectories** (called "CCA"- Complete Case Analysis) without introducing bias into our analysis.

We use Bayesian multiple imputation to model the missing data given the observed data. This technique uses Markov Chain Monte Carlo sampling with the General Location Model and non-informative priors to produce a collection of *imputed* or "filled-in" training sets. **A Q-function is learned for each imputed training set, then these Q-functions are averaged to produce a single expected Q-function, with the expectation taken over the missing data.** This reduces non-response bias.

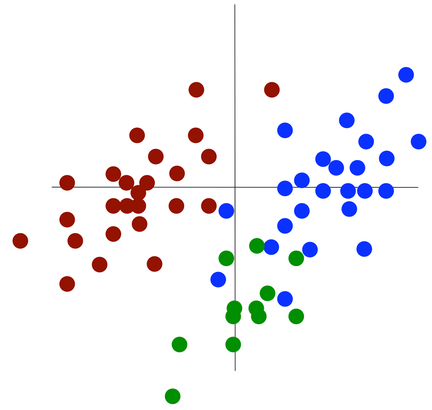
Finally, **we use the non-parametric bootstrap to assess the confidence of the resulting model**. We generate 1000 bootstrapped incomplete training sets, impute them as described above, and determine how many of the 1000 resulting Q-functions prefer each action. These votes are shown in the bar graphs to the right.

We acknowledge support for this project from National Institute of Health grants RO1 MH080015 and P50 DA10075. We also thank the STAR*D research team (http://www.star-d.org/).

The Model

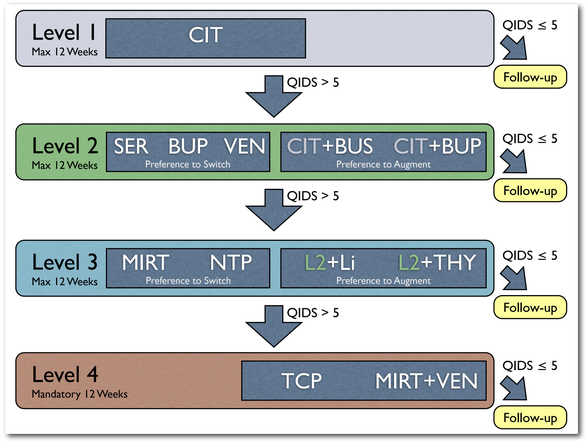
We use the "General Location Model." Discrete variables are modeled as a multinomial, and continuous variables are modeled as a mixture of Gaussians with one component for each discrete variable configuration. Covariance is shared among all the components.

$$
\begin{aligned}
X &= [W, Z_1, Z_2] \\
W &\in \{1, 2, 3\} \\
[Z_1, Z_2] &\in \mathbb{R}^2 \\
P(W) &= [0.4, 0.4, 0.2] \\
\mu(1) &= [-1.0, 0.0] \\
\mu(2) &= [1.0, 0.0] \\
\mu(3) &= [0.0, -1.0] \\
\Sigma &= \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}
\end{aligned}
$$



The Data

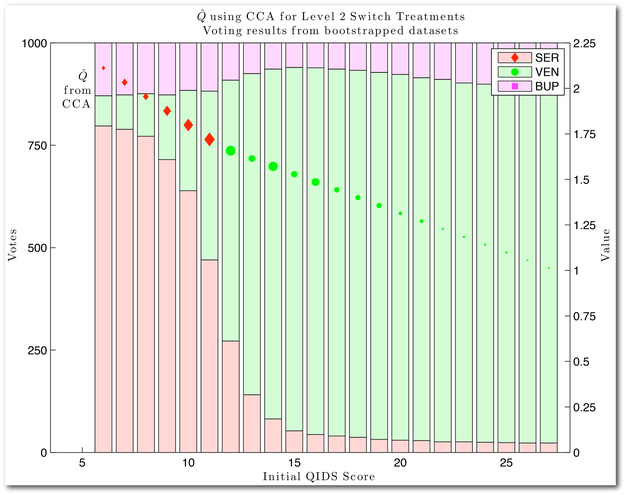
In the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study, patients followed a depth-four randomized treatment policy.



The Results

The STAR*D patient trajectories are analyzed using Q-learning with linear function approximation. Confidence is shown in bar graphs of votes from 1000 bootstrapped training sets. Complete Case Analysis (CCA) produces high confidence in SER as the optimal Level 2 treatment for low QIDS-SR_{16} patients, and VEN for high QIDS-SR_{16} patients. This is due in part to non-response bias. Analysis using Bayesian multiple imputation is more conservative, and better represents the total sample.

Complete Case Analysis



Multiple Imputation