Variable selection and machine learning methods in causal inference

Debashis Ghosh

Department of Biostatistics and Informatics
Colorado School of Public Health
Joint work with Yeying Zhu, University of Waterloo and Wei Luo, Baruch College, CUNY

UC Denver STARS Seminar
Outline

1. Introduction
2. Potential Outcomes Framework
3. LASSO algorithms for Causal Inference
4. Application
5. Causal Inference and Balance
6. Dimension Reduction for Causal Inference
7. Discussion
Causal inference has recently received intense interest in biomedical studies.

While it is motivated by observational data, there has also been recent interest in the use of these models in clinical trials to understand things like surrogate endpoints (Li et al., 2009, 2010; Ghosh et al., 2010).
In these settings, the treatment is not randomly assigned and is subject to self-selection/confounding.

A very important modelling strategy was proposed 30 years ago by Rosenbaum and Rubin (1983), termed the propensity score.

In words, the propensity score is defined as the probability of receiving treatment, given covariates.

Conditional on propensity score, one achieves “covariate balance” on observed covariates.
Potential Outcomes: Notation and Assumptions

- Let $T \in \{0, 1\}$ denote the treatment
- Let $\{Y(0), Y(1)\}$ denote the potential outcomes for $Y$ under each of the treatments
- Standard assumption necessary for causal inference:

$$T \perp \{Y(0), Y(1)\} | X$$

where $X$ are covariates
Potential Outcomes (cont’d.)

 Targets of estimation:

\[
ACE = n^{-1} \sum_{i=1}^{n} \{ Y_i(1) - Y_i(0) \}
\]

and

\[
ACET = n^{-1} \sum_{i=1}^{n} \{ Y_i(1) - Y_i(0) \} \times I(T_i = 1)
\]
Propensity score: $P(T = 1|X)$

- Estimate propensity score using logistic regression
- Given estimated propensity scores, one can estimate ACE and ACET in a variety of ways
  1. Matching
  2. Inverse probability weighted estimating equations
  3. Regression using propensity score as covariate
What variables to include in propensity score model?

- Advice from Rosenbaum and Rubin: include everything associated with outcome $Y$
- Advice from Pearl: think about a graph structure for relating the variables, do NOT include “colliders”
- Many simulation studies reported on in the literature.
- Recent interest in Bayesian model averaging and related approaches for this problem (Zigler and Dominici, 2013; Wang et al., 2012, 2015; Polley et al., 2007)
Propensity score modelling: remarks

- We do not seek to interpret the model for the propensity score
- We only use the estimated probabilities from the fitted model at the second stage
- The real target of estimation is the causal effect
Models and variable selection

- Two models: Propensity score model and mean outcome model
- Problems with applying ‘off-the-shelf’ variable selection procedures to either model for performing variable selection
- Mean outcome model: applying variable selection would change causal scientific estimand
- For example, ACE corresponds to

\[ Y(1) - Y(0) = \tau + \epsilon, \]

which is different from

\[ Y(1) - Y(0) = \tau^* + \delta'X + \epsilon^* \]
Example of why variable selection for propensity score model is not sufficient
### Causal Inference as Missing Data Problem

- **Data visualization**

<table>
<thead>
<tr>
<th>$Y(0)$</th>
<th>$Y(1)$</th>
<th>$X_1$</th>
<th>$\cdots$</th>
<th>$X_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ ? $</td>
<td>$y_1$</td>
<td>$x_{11}$</td>
<td>$\cdots$</td>
<td>$x_{1p}$</td>
</tr>
<tr>
<td>$y_2$</td>
<td>$\ ? $</td>
<td>$x_{21}$</td>
<td>$\cdots$</td>
<td>$x_{2p}$</td>
</tr>
<tr>
<td>$\ ? $</td>
<td>$y_3$</td>
<td>$x_{31}$</td>
<td>$\cdots$</td>
<td>$x_{3p}$</td>
</tr>
<tr>
<td>$\ ? $</td>
<td>$y_4$</td>
<td>$x_{41}$</td>
<td>$\cdots$</td>
<td>$x_{4p}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\cdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$y_n$</td>
<td>$\ ? $</td>
<td>$x_{n1}$</td>
<td>$\cdots$</td>
<td>$x_{np}$</td>
</tr>
</tbody>
</table>
The missing data mechanism and SITA assumption

\[ \{ Y(0), Y(1) \} \perp T | X \]

suggests the following two-step algorithm

1. Fill in missing responses (imputation)
2. Perform variable selection on the ‘complete’ data

Note that \( \{ Y(0), Y(1) \} \) is a multivariate response variable, so variable selection becomes dependent on the joint distribution of the potential outcomes.
Introduction
Potential Outcomes Framework
LASSO algorithms for Causal Inference
Application
Causal Inference and Balance
Dimension Reduction for Causal Inference
Discussion

Difference LASSO algorithm

1. Fit a regression model for $Y$ on $X$ for individuals with $T = 0$ and $T = 1$ separately. This will yield two prediction models, one for the treated group ($T = 1$), and one for the control group ($T = 0$). We will denote these as models $M_0$ and $M_1$.

2. Based on the models fitted in step 1., compute predicted/fitted values of $Y$ using the test dataset to impute the counterfactual or potential outcome described in Section 2. In particular, we will use $M_0$ to predict $Y(0)$ for subjects with $T = 1$ and $M_1$ to impute $Y(1)$ for subjects with $T = 0$.

3. Compute the average of the potential outcomes, $Y^d$, and perform a LASSO of $Y^d$ on $X$.
LASSO: remarks

- Both are applications of the *predictive* LASSO idea of Tran et al. (2012)
- Use cyclic coordinate descent algorithm of Friedman et al. (2010)
- Apply LASSO to functionals of the predictive distribution, not just to a regular observed data likelihood.
- This will complicate inference (open topic)
- The difference LASSO identifies variables that define subgroups for which the average causal effect is homogeneous.
LASSO: remarks (cont’d.)

- Can be motivated from a hierarchical model structure similar to one used in George and McCulloch (1993)
- Link with missing data and imputation methods
- The previously proposed algorithms correspond to “single” imputation estimators
- Can also derive multiple imputation procedures, termed multiple difference LASSO.
Illustration

- Data from Connors et al. (1996)
- Goal: study effect of right-heart catherization on survival
- Response: 30-day survival
- 74 variables in the dataset, but here, we consider 21 variables
Illustration (cont.’d)
Illustration (cont.’d)
Causal inference: a revisit

- Dominant model in the field: potential outcomes framework
- Let $Y$ be response of interest, $T$ be nonrandomized binary treatment and $X$ be the confounders.
- Multi-stage modelling process:
  1. Model the propensity score $P(T|X)$
  2. Match on the propensity score
  3. Check for balance between treatment groups in matched sample
  4. Estimate average causal effect
- Our focus is on step 3).
- One state of the art balance approach: genetic algorithms (Sekhon), very computationally expensive.
Covariate Balance

- Fundamentally a comparison of $X|T = 0$ and $X|T = 1$
- Theory: equal percent bias reduction of Rubin and collaborators
- In practice: two-sample $t$-tests typically are done before and after matching
- Recent innovation: CBPS of Imai and Ratkovic (2014) for propensity scores that achieve balance.
Our proposal: probability metrics

- Use

\[ \gamma(P, Q) = \sup_{f \in \mathcal{F}} \left| \int f dP - \int f dQ \right|, \tag{1} \]

where \( \mathcal{F} \) is a class of functions, to assess balance between \( P \) (probability law for \( X \mid T = 0 \)) and \( Q \) (probability law for \( X \mid T = 1 \))

- If \( \gamma(P, Q) = 0 \), then \( P \) and \( Q \) induce equivalent probability spaces

- If \( \mathcal{F} \) corresponds to an RKHS, then \( \gamma(P, Q) \) will have a simple empirical estimator that is evaluable in closed form.

- We term (8) the kernel distance and use as our statistic for evaluating balance
Some simulation studies

- Simulate nine covariates, mix of continuous and binary.
- treatment variable $T$ is generated from $\text{Bernoulli}(e(Z))$
  where
  \[
  \text{logit}(e(Z)) = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \alpha_3 Z_4 + \alpha_4 Z_5 + \alpha_5 Z_7 + \alpha_6 Z_8 + \alpha_7 Z_2 Z_4 \\
  + \alpha_8 Z_2 Z_7 + \alpha_9 Z_7 Z_8 + \alpha_{10} Z_4 Z_5 + \alpha_{11} Z_1 Z_1 + \alpha_{12} Z_7 Z_7,
  \]
  and
  \[
  \alpha = (0, \log(2), \log(1.4), \log(2), \log(1.4), \log(2), \log(1.4), \log(1.2), \log(1.4), \log(1.6), \log(1.2), \log(1.4), \log(1.6)).
  \]
- The outcome variable $Y$ is generated from four different scenarios (A,B,C,D) which differ in terms of model complexity (taken from Stuart et al. (2013))
- True causal effect is a constant: $\gamma = 3$. 
Some simulation studies (cont’d.)

- Fit the same types of propensity score models as in Stuart et al. (2013), many of which will be misspecified.
- 1-1 matching, focus on the average causal effect of the treated (ACET).
- Misspecified propensity score $\rightarrow$ imbalance in covariates $\rightarrow$ biased causal effects.
- Goal of balance statistics is to detect this imbalance.
- Metric of evaluation: correlation between balance statistic and bias.
- Comparisons with average standardized mean difference (ASMD)-based balance statistics and Kolmogorov-Smirnov (KS).
### Table: Mean and Standard Deviation of the Pearson Correlation Coefficients

<table>
<thead>
<tr>
<th>Matching: Mean (SD)</th>
<th>Outcome A</th>
<th>Outcome B</th>
<th>Outcome C</th>
<th>Outcome D</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean ASMD</td>
<td>0.632 (0.134)</td>
<td>0.609 (0.155)</td>
<td>0.606 (0.152)</td>
<td>0.587 (0.156)</td>
</tr>
<tr>
<td>max ASMD</td>
<td>0.557 (0.176)</td>
<td>0.542 (0.196)</td>
<td>0.548 (0.185)</td>
<td>0.512 (0.193)</td>
</tr>
<tr>
<td>median ASMD</td>
<td>0.367 (0.214)</td>
<td>0.351 (0.212)</td>
<td>0.347 (0.221)</td>
<td>0.356 (0.213)</td>
</tr>
<tr>
<td>mean KS</td>
<td>0.372 (0.264)</td>
<td>0.358 (0.267)</td>
<td>0.348 (0.276)</td>
<td>0.333 (0.273)</td>
</tr>
<tr>
<td>mean t-statistic</td>
<td>0.634 (0.133)</td>
<td>0.609 (0.155)</td>
<td>0.610 (0.149)</td>
<td>0.588 (0.155)</td>
</tr>
<tr>
<td>kernel distance</td>
<td>0.797 (0.115)</td>
<td>0.773 (0.140)</td>
<td>0.788 (0.120)</td>
<td>0.759 (0.125)</td>
</tr>
</tbody>
</table>
Other innovations

- Have developed a version of genetic matching with kernel distance as the balance metric.
- Intuition: Balance is about constraining moments, i.e.,
  \[ \text{mean}(X_1|T = 1) = \text{mean}(X_1|T = 0) \]
- Our kernel constrains the *functional forms* of variables simultaneously, i.e., we want
  \[ f(X|T = 1) = f(X|T = 0) \]
  to hold simultaneously for as many functions as possible.
- The Gaussian kernel corresponds to a very rich class of functions (dense in \(L_2(\mathcal{X})\))
Conditional independence assumptions for causal inference

- Recall SITA assumption from causal inference intro:
  \[ T \perp \{ Y(0), Y(1) \} | X \]
  where \( X \) are covariates
- For estimating average causal effect (ACE), this can be relaxed to
  \[ T \perp Y(0) | X \]
  and
  \[ T \perp Y(1) | X \]
Conditional independence assumptions for causal inference (cont’d.)

• Suppose we add an assumption from the literature on dimension reduction (Li, 1991; Cook, 1998):

\[ E(Y(0) | X) \perp X | X' \beta_0 \]

and

\[ E(Y(1) | X) \perp X | X' \beta_1 \]

where \( \beta_0 \) is \( p \times r(0) \) and \( \beta_1 \) is \( p \times r(1) \)

• These are called central mean subspaces, and \( r(0) \) and \( r(1) \) are the dimensions
Key results

- Under central mean subspace assumption and SITA, one can estimate the directions of the mean potential outcomes (conditional on covariates) using the observed data.
- This lends itself to a natural dimension reduced-based algorithm for causal inference.

1. Estimate $Y(1)$ and $Y(0)$ from the observed data using a dimension reduction method (we use MAVE from Xia et al., 2002).
2. Compute the difference and average over covariates to get an average causal effect estimate.
Key results (cont’d.)

- Very similar in spirit to G-computation algorithm of Jamie Robins
- Implicitly, the propensity score is estimated
- This approach allows for the overlap assumption to be relaxed
- If \( r(0) \) and \( r(1) \) are relatively smallled compared to \( p \), we can achieve superefficiency (like in LASSO)
Causal inference poses a very interesting model selection problem

Adopting the potential outcomes and missing data framework clarifies difficulty in variable selection

“Impute/Penalize" algorithm is fairly general

Role of prediction

Open issue: Post-model selection inference/standard errors
References


- Contact [debashis.ghosh@ucdenver.edu](mailto:debashis.ghosh@ucdenver.edu) for last two papers.