Propensity Score Estimation with Machine Learning Methods on Electronic Health Databases with Healthcare Claims Data

Cheng Ju

Division of Biostatistics, University of California, Berkeley

under supervision of Professor Mark van der Laan

Qualifying Exam Committee (alphabetical order):
John F. Canny
Alan Hubbard (Chair)
Nicholas P. Jewell
Mark J. van der Laan

Propensity Score

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- The propensity score is the probability of a unit (e.g., person, classroom, school) being assigned to a particular treatment given a set of observed covariates.
- Propensity scores can be used to reduce selection bias.
- This project mainly focus on the estimation of PS on electronic health databases with claims data.

Motivation

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Healthcare claims data can be understood and analyzed as a set of proxies that indirectly describe the health status of patients.
- Adjusting for a surrogate of an unmeasured factor usually helps to adjusting for the factor itself.
- The patients healthcare claims data may improve the bias reduction in the observational study.

Example of Claims Data

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Each claims code covariate records the number of times a claims code occurred for each patient.
- A patient has a value of 2 for the variable pxop V5260, then the patient received twice the outpatient procedure, which is coded as V5260.

	pxop V5260
observation 1	0
observation 2	2
observation 3	1
⋮	⋮

Challenges

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- The claims data is usually high-dimensional, with large amount of covariates.
- Example: In later data analysis, the data in NOAC study has 18,447 observations, 60 baseline covariates (e.g. gender, height, weight) and **23,531** claims code covariates.
- The claims code data is highly sparse: there is usually few non-zero value code for each patient (row).
- This leads to two challenges: 1. low signal-to-noise ratio; 2. curse of dimensionality.
- It is hard for human expert to check manually which variable is potential confounder.

High-dimensional Propensity Score Methods

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

Simple rule-based feature generating and screeing procedure. High-dimensional Propensity Score Methods [1] includes 7 steps:

1. Manually Identify data sources (dimensions).
2. Empirically identify candidate covariates.
3. Assess recurrence of codes.
4. Prioritize covariates.
5. Select covariates for adjustment.
6. Estimate the exposure propensity score.
7. Make inference on the target parameter.

[1] Schneeweiss, Sebastian, et al. "High-dimensional propensity score adjustment in studies of treatment effects using health care claims data." Epidemiology (Cambridge, Mass.) 20.4 (2009): 512. APA

High-dimensional Propensity Score Methods

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

The hdPS algorithm has two tuning-parameter (hyper-parameter), n and k. n is the number of covariate kept within each source, and k is the total number of covariatets kept during the screening process. More details later.

High-dimensional Propensity Score Methods

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Step 1: Manually Identify data sources (dimensions). For example, cluster claims data into diagnoses codes, drug claims, and procedure codes.

High-dimensional Propensity Score Methods

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Step 2: Empirically identify candidate covariates.
  For each data source (dimension), **sort the codes by their empirical prevalence and select top n covatiets within each source**. For code covariates x, the prevelance is defined as $\max(p_{n,x}, 1 - p_{n,x})$, where $p_{n,x}$ is the proportion of observations having non-zero value for this code.
  Assume we have J sources in the last step, then we have $J \times n$ claims covariates left after this step.

High-dimensional Propensity Score Methods

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Step 3: Assess recurrence of codes.
  For each identified code, create three indicator covariates: "once" for appearing at least once, "sporadic" for appearing more than the median, and "frequent" for appearing more than the 75th percentile.
  Example: Consider for code "pxop V5260" we mentioned before. We generate 3 indicator variable for this code. Assume it has median $c_{0.5}$ and 75% quantile $c_{0.75}$
  For i-th observation, generate three covariates:
  pxop V5260 once: $I(\text{pxop V5260} > 0)$
  pxop V5260 sporadic: $I(\text{pxop V5260} > c_{0.5})$
  pxop V5260 frequent: $I(\text{pxop V5260} > c_{0.75})$

High-dimensional Propensity Score Methods

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- We then remove the original claims covariates. Thus we have $J \times n \times 3$ columns left in total.
- For simplicity, we call them hdPS covariates.

High-dimensional Propensity Score Methods

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Step 4: Prioritize covariates. For each hdPS covariate from last step, estimate the potential confounding impact $|\log(Bias_M)|$ for the variable C using the Bross formula[2] :

$$Bias_M = \frac{P_{C1}(RR_{CD} - 1) + 1}{P_{C0}(RR_{CD} - 1) + 1}$$

where $P_{C1} = P_n(C = 1|A = 1)$, $P_{C0} = P_n(C = 1|A = 0)$, and $RR_{CD} = \frac{P_n(Y=1|C=1)}{P_n(Y=1|C=0)}$ is the risk-ratio for outcome Y with different level of C.

[2]The formula in hdPS paper (Schneeweiss et al. 2009) is different from Bross 1966. Here we followed Bross paper.

Note: Step 4 uses the treatment and outcome covariates to rank the claims covariates. In the later experiments, we only use the training set to 'prioritize' the covariates to make sure the testing data is remain untouched.

High-dimensional Propensity Score Methods

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Step 5: Select covariates for adjustment: select the top k empirically ranked covariates from step 4. Consider in step 2, we keep n covariates for all J sources (dimensions). After step 3, we would have $3 \times n \times J$ hdPS covariates. In this step, we select k among these features by the ranking from step 4.

High-dimensional Propensity Score Methods

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Step 6: Estimate the exposure propensity score. Using multivariate logistic regression, a propensity score was estimated for each subject as the predicted probability of exposure conditional on both the baseline covariates and hdPS covariates.
- Step 7: Make inference based on the estimated PS. This step we could simply apply any models that rely on the propensity score, like inverse probability of treatment weighted estimator (IPTW).

Defining the Estimation Problem

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Data structure: $O = (X, Y) \sim P_0$ and we observe n **i.i.d.** observations on $O_1 \cdots O_n$.
- Parameter of interest: $\psi_0(X)$ defined as minimizer of a loss function $L(O, \Psi)$, over a parameter space Ψ:

$$\psi_0 = \arg\min_{\psi \in \Psi} \mathbb{E}_0(L(O, \psi))$$

- Example: L2 loss for the regression problem:

$$L(O, \Psi) = (Y - \Psi(X))^2$$

$$\Psi_0 = E_0(Y|X)$$

Motivation of Adaptive Ensemble Learning

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- More and more choices for such estimation problem...
- In practice, it is generally impossible to know a priori which learner will perform best for a given prediction problem and data set.

Discrete Super Learner

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

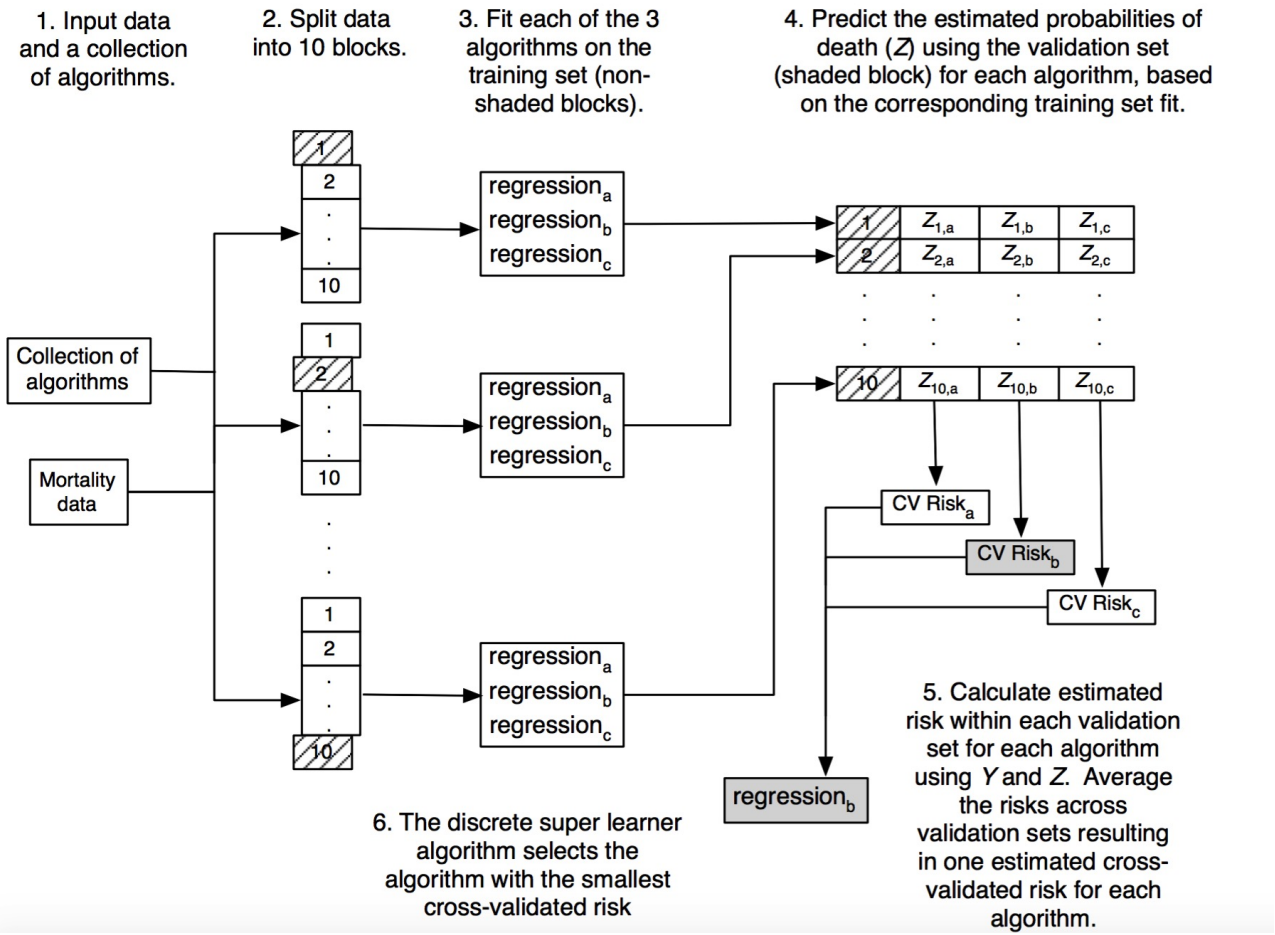


Figure: The procedure of discrete Super Learner from chapter 3 of van der Laan and Rose 2011.

Oracle Inequalities

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- If, as is typical, none of the candidate learners (nor, as a result, the oracle selector) converge at a parametric rate, the super learner performs asymptotically as well (in the risk difference sense) as the oracle selector, which chooses the best of the candidate learners.
- If one of the candidate learners searches within a parametric model and that parametric model contains the truth, and thus achieves a parametric rate of convergence, then the super learner achieves the almost parametric rate of convergence $\log(n)/n$.

- Discrete SL is a selector. It only uses one 'best' algorithm w.r.t cross-validated loss.
- Consider a set of all the possible weights for the base learners. It is reasonable to expect that one of these weighted averages might perform better than one of algorithm alone.
- Could use discrete SL to select the best algorithm in a library, which consists all the possible convex combination of the base learners.
- This is the Super Leaner.

Super Learner

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

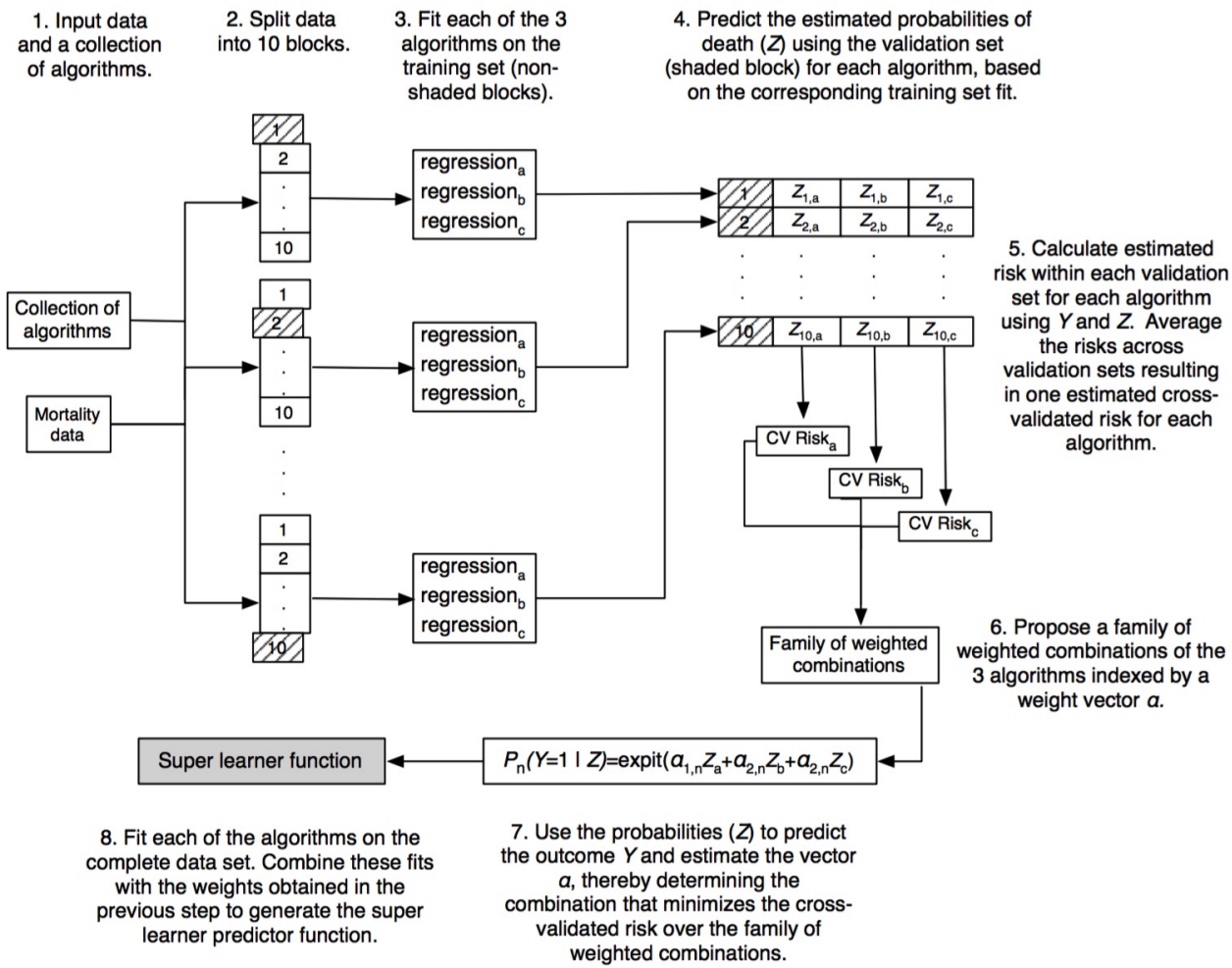


Figure: The procedure of Super Learner from from chapter 3 of van der Laan and Rose 2011.

Related Works

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- SL has been shown to perform well in a number of settings: Sinisi et al. 2007 used SL to predict HIV-1 drug resistance; Gruber et al. 2015 used SL to estimate the inverse probability weights for marginal structural modeling in large observational datasets; Rose 2016 applied SL for Plan Payment Risk Adjustment.
- However, the performance of SL has not been thoroughly investigated within large electronic healthcare datasets based on healthcare claims data, that are becoming common in medical research.

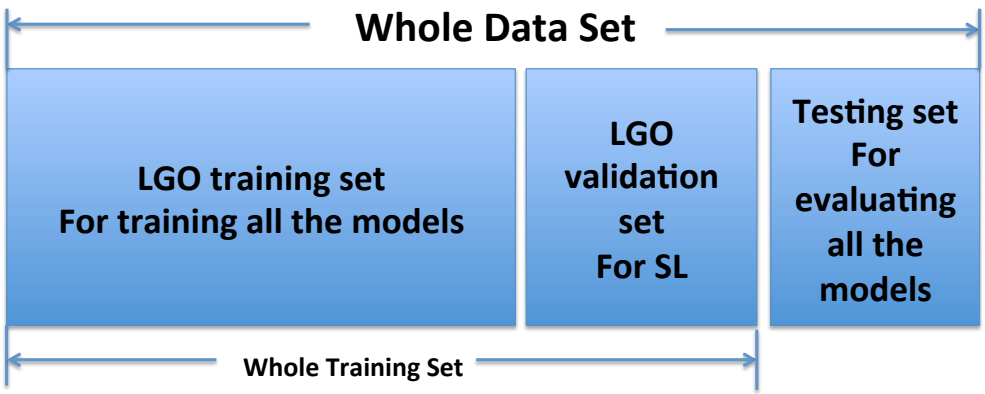
Sample Split Super Learner

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

Super Learner computes the weight by minimizing the cross-validated risk. When data is too large, leave one group out (LGO) validation could be used to reduce the computation.

1. Models are firstly trained on LGO training set.
2. The SL weight is computed on LGO validation set.
3. Finally all the models are re-trained on whole training set.

Performance of Sample-Split Super Learner is assessed on a untouched testing set.



Machine Learning Library for Super Learner

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- We used 23 statistcal/machine learning algorithms based on caret R package API.
- The library covers almost all the popular methods, like Boosting, glm, and glmnet. Some algorithm with high time complexity (e.g. random forest) are removed.

Machine Learning Library for Super Learner

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

Bayesian Generalized Linear Model ("bayesglm"),
C5.0 ("C5.0"),
Single C5.0 Ruleset ("C5.0Rules"),
Single C5.0 Tree ("C5.0Tree"),
Conditional Inference Tree, ("ctree")
Multivariate Adaptive Regression Spline ("earth"),
Boosted Generalized Linear Model ("glmboost"),
Penalized Discriminant Analysis ("pda"),
Shrinkage Discriminant Analysis ("sda"),
Flexible Discriminant Analysis ("fda"),
Lasso and Elastic-Net Regularized Generalized Linear Models ("glmnet"),
Penalized Discriminant Analysis ("pda2"),
Stepwise Diagonal Linear Discriminant Analysis ("sddaLDA"),
Stochastic Gradient Boosting ("gbm"),
Multivariate Adaptive Regression Splines ("gcvEarth"),
Boosted Logistic Regression ("LogitBoost"),
Penalized Multinomial Regression ("multinom"),
Penalized Logistic Regression ("plr"),
CART ("rpart"),
Stepwise Diagonal Quadratic Discriminant Analysis ("sddaQDA"),
Generalized Linear Model ("glm"),
Nearest Shrunken Centroids ("pam"),

Cost-Sensitive CART ("rpartCost")

Super Learners

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

For simplicity, we introduce two Super Learners:

Name	Library	Covariates
SL1	All machine learning algorithms	Only baseline covariates.
SL2	All machine learning algorithms and the hdPS algorithm	Baseline covariates; Only the hdPS algorithm can use claims data.

Table: Details of the two Super Learners considered.

Novel Oral Anticoagulant (NOAC) Study

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- The data was collected by United Healthcare, recorded between October, 2009 and December, 2012.
- The dataset includes 18,447 observations, 60 baseline covariates and 23,531 claims code covariates.
- **Outcome variable:** one for patients who had a stroke and zero for the others.
- **The exposure:** warfarin (0) or dabigatran (1).
- **Claims code covariate:** The claims code covariates fall into four sources, or "data dimensions": inpatient diagnoses, outpatient diagnoses, inpatient procedures, and outpatient procedures.

Nonsteroidal anti-inflammatory drugs (NSAID) Study

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- The observations were sampled from a population of patients aged 65 years and older enrolled in both Medicare and the Pennsylvania Pharmaceutical Assistance Contract for the Elderly (PACE) programs between 1995 and 2002.
- There were 49,653 observations, with 22 baseline covariates and 9,470 claims code covariates in this study.

Nonsteroidal anti-inflammatory drugs (NSAID) Study

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- **Outcome variable:** 1 for severe gastrointestinal (GI) complication and 0 for the others.
- **The exposure:** 1 for Cox-2 inhibitors on reduced gastric toxicity, and 0 for nonselective nonsteroidal anti-inflammatory drugs.
- **Claims code covariate:** The claims code covariates fell into eight data sources: prescription drugs, ambulatory diagnoses, hospital diagnoses, nursing home diagnoses, ambulatory procedures, hospital procedures, doctor diagnoses and doctor procedures.

Vytorin Study

- The observation is all United Healthcare patients linked for January 1, 2003 to December 31, 2012, with age over 65 on day of entry into cohort.
- There were $148,327$ observations, with 67 baseline covariates and $15,010$ claims code covariates in this study.
- **Outcome variable:** 1 for myocardial infarction, stroke, or death. 0 for the others.
- **The exposure:** 1 for Vytorin and 0 for high-intensity statin therapies.
- **Claims code covariate:** The claims code covariates fall into five sources: ambulatory diagnoses, ambulatory procedures, prescription drugs, hospital diagnoses and hospital procedures.

Performance Metrics

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Two class (treatment and control) are usually not well balanced. Prediction accuracy is not a good metric to assess the models.
- For VYTORIN data, only 16% are in the treatment group, so naive classifier that predicting all the propensity score with 0 would have around 84% accuracy.
- In the inference step, knowing the most possible label is not sufficient: **we need a good estimate of propensity score (probability), instead of just the most possible label.**

Performance Metrics

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Negative log-likelihood (cross-entropy) is commonly used for unbalanced classification. Negative log-likelihood is used as loss function for Super Learner to compute the optimal weight.
- AUCROC curve: draw the ROC (receiver operating characteristic) curve, which is the true positive rate against the false positive rate at various threshold settings. Then compute the area under the ROC curve.

Running Time

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

Data Set	Algorithm	Processing Time (seconds)
NOAC	Sum of ML algorithms	481.13
	Sum of hdPS algorithms	222.87
	Super Learner 1	1035.43
	Super Learner 2	1636.48
NSAID	Sum of ML algorithms	476.09
	Sum of hdPS algorithms	477.32
	Super Learner 1	1101.84
	Super Learner 2	2075.05
VYTORIN	Sum of ML algorithms	3982.03
	Sum of hdPS algorithms	1398.01
	Super Learner 1	9165.93
	Super Learner 2	15743.89

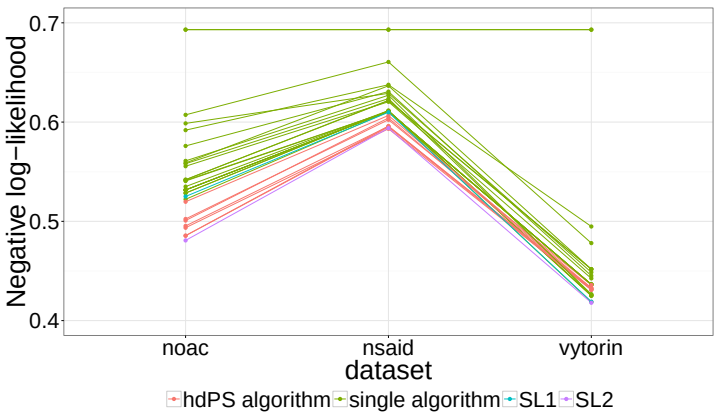
Table: Running time for all algorithms.

The models are trained on BWH cluster with single CPU[3].

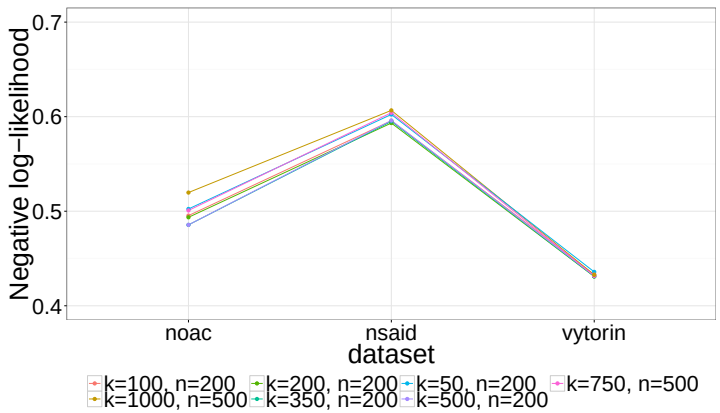
[3]Intel Xeon CPU E7- 4850, 2.00GHz

Negative log-likelihood

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix



(a) Negative log-likelihood for SL1, SL2, the hdPS algorithm, and the 23 machine learnng algorithms.

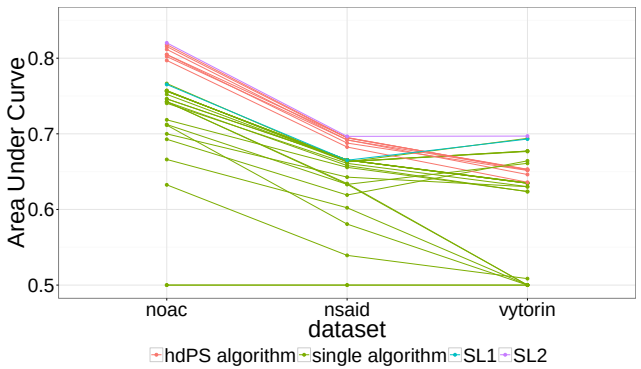


(b) Negative log-likelihood for the hdPS algorithm, varying the parameter k from 50 to 750 for $n = 200$, and $n = 500$.

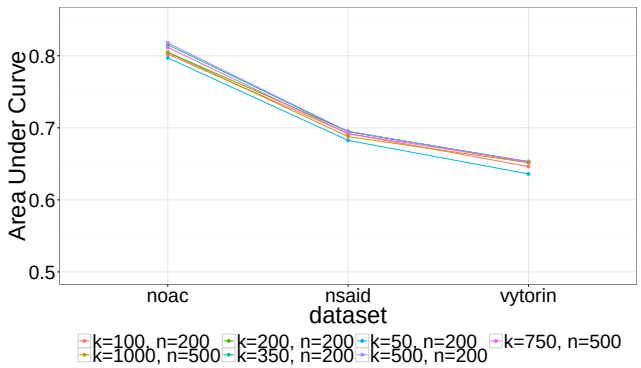
Figure: The negative log-likelihood for SL1, SL2, the hdPS algorithm, and the 23 machine learning algorithms. SL 1 outperforms all conventional ML algorithms, and have similar performance to hdPS algorithms. SL2 outperforms all algorithms.

AUCROC

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix



(a) AUC of SL1, SL2, the hdPS algorithm, and the 23 machine learnng algorithms.



(b) AUC for the hdPS algorithm, varying the parameter k from 50 to 750 for $n = 200$, and $n = 500$.

Figure: The area under the ROC curve (AUC) for for Super Learners 1 and 2, the hdPS algorithm, and each of the 23 machine learning algorithms. Result is similar to negtive log-likelihood

Weight of the Algorithms in the Super Learners

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

Data Set	SL1	Weight	SL2	Weight
NOAC	C5.0	0.11	earth	0.05
	bayesglm	0.30	hdps.500_200	0.19
	gbm	0.39	hdps.350_200	0.48
NSAID	C5.0	0.06	hdps.1000_500	0.23
	glm	0.35	gbm	0.24
	gbm	0.52	hdps.100_200	0.25
VYTORIN			hdps.350_200	0.07
	multinom	0.07	hdps.1000_500	0.17
	gbm	0.93	gbm	0.71

Table: Base Learners with top 3 weights in Super Learners 1 and 2 across all three data sets. hdPS methods dominated Super Learner except VYTORIN data set.

- Algorithm with best performance does not guarantee to have highest weight: SDA have similar performance with gbm, but weight is 0.
- In addition, even the relative weights changed a lot after adding new algorithm into the library.

Rethinking of hdPS Algorithm

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Original hdPS method generates and screens the hdPS covariates, and uses **logistic regression** to predict the propensity score.
- Why not changing the prediction algorithm from logistic regression, to other supervised learning algorithm?
- Instead including "hdPS prediction algorithm" in the library, we simply provide "hdPS covariates" to all the algorithms!

Rethinking of hdPS Algorithm

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- For simplicity, we denote **"hdPS prediction algorithm"** as the step 1-6 of the full hdPS algorithm in the original paper (excluding the inference step), as here we mainly focus on the prediction task. In addition, we denote **"hdPS screening method"** as the step 1-5, which could be considered as a feature generating/selecting algorithm.

hdPS Features

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Pros: More flexible choice of prediction algorithms: assumptions of logistic regression is usually not reasonable, as the data generating system is usually non-linear.
- Cons: Less flexible for hdPS parameter selection: it takes too much time if put **all the combinations** of ML algorithms and hdPS covariates that generated by different tuning parameter (k, n).

hdPS Features

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Fix the hyper-parameter for hdPS screening method: use cross-validation on the **LGO-training set** to "pre-select" the tuning parameter of hdPS, w.r.t. its predictive performance.
- Then only use selected hyper-parameter pair to generate hdPS features for all the algorithms in the library of Super Learner.

Three Super Learners

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

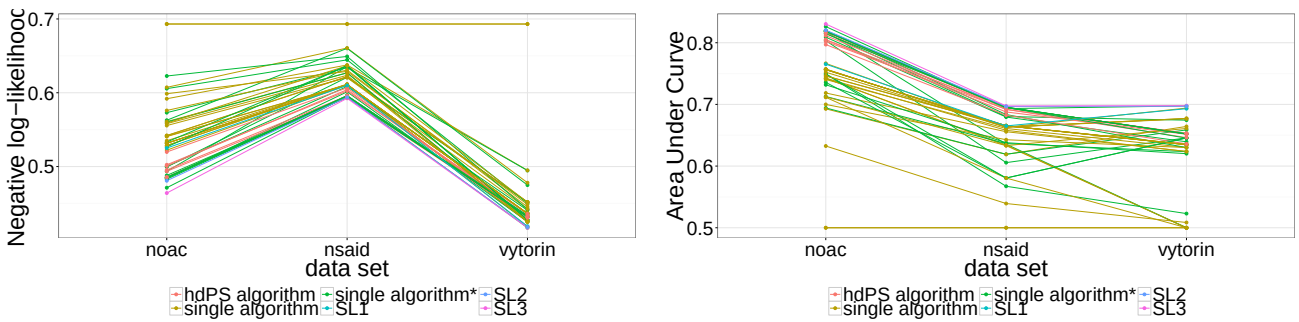
Name	Library	Covariates
SL1	All machine learning algorithms	Only baseline covariates.
SL2	All machine learning algorithms and the hdPS algorithm	Baseline covariates; Only the hdPS algorithm can use the claims data.
SL3	All machine learning algorithms	Baseline covariates and hdPS covariates generated from the claims data by hdPS screening method.

Table: Details of the three Super Learners considered.

Note the origianl hdPS algorithm uses logistic regression for prediction, which is a special case of glmnet.

Performance

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix



(a) Negative log-likelihood (b) AUC

Figure: Negative log-likelihood and AUC of SL1, SL2, and SL3, compared with each of the single machine learning algorithms (with and without using hdPS covariates). We could see among all the single algorithms and Super Learners, SL3 performs the best cross three datasets. More details in the next table.

Comparison of Super Learners

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

Data set	Metric	best ML	best hdPS	best ML with hdPS	SL 1	SL 2	SL 3
NOAC		0.766(gbm)	0.818	0.826(gbm)	0.7652	0.8203	0.8304
NSAID	AUC	0.666(sda)	0.695	0.696 (sda)	0.6651	0.6967	0.6975
VYTORIN		0.694(gbm)	0.653	0.698 (gbm)	0.6931	0.6970	0.6980
NOAC		0.522(gbm)	0.486	0.471 (gbm)	0.5251	0.4808	0.4641
NSAID	NLL	0.610 (gbm)	0.594	0.594 (sda)	0.6099	0.5939	0.5924
VYTORIN		0.420 (gbm)	0.431	0.418 (gbm)	0.4191	0.4180	0.4170

Table: Performance as measured by AUC and negative log-likelihood for the three Super Learners.

- Difference between SL1 and SL2 is large (for NOAC and NSAID data), while SL2 and SL3 is mild.
- Extra features from claims code contribute a lot to the prediction of propensity score for NOAC and NSAID data.

Comparison of Super Learners

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

Data set	Metric	best ML	best hdPS	best ML with hdPS	SL 1	SL 2	SL 3
NOAC		0.766(gbm)	0.818	0.826(gbm)	0.7652	0.8203	0.8304
NSAID	AUC	0.666(sda)	0.695	0.696 (sda)	0.6651	0.6967	0.6975
VYTORIN		0.694(gbm)	0.653	0.698 (gbm)	0.6931	0.6970	0.6980
NOAC		0.522(gbm)	0.486	0.471 (gbm)	0.5251	0.4808	0.4641
NSAID	NLL	0.610 (gbm)	0.594	0.594 (sda)	0.6099	0.5939	0.5924
VYTORIN		0.420 (gbm)	0.431	0.418 (gbm)	0.4191	0.4180	0.4170

Table: Performance as measured by AUC and negative log-likelihood for the three Super Learners.

- More flexible ML algorithms improved the performance, but not much. This suggests multivariate logistic regression already have satisfactory performance, in these data sets.
- For Vytorin, the performance of the best ML and the best ML with hdPS is close. This suggests the claims data does not contain extra helpful information to predict treatment mechanism in this data set.

Sensitivity of hyper-parameter of hdPS algorithm

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

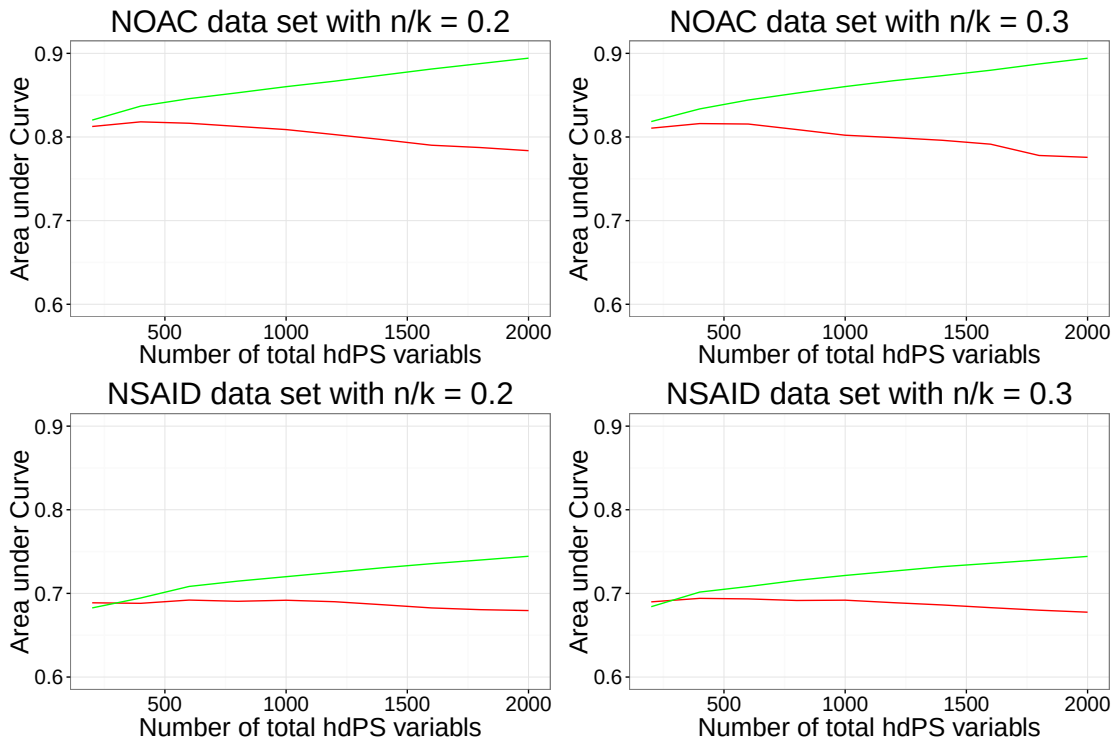


Figure: AUC for hdPS algorithms with different number of variables, k. We fixed n/k and increased k to investigate the change of AUC. The AUC for prediction is not sensitive for the total number of hdPS covariates selected, k.

Further study of hdPS algorithm

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

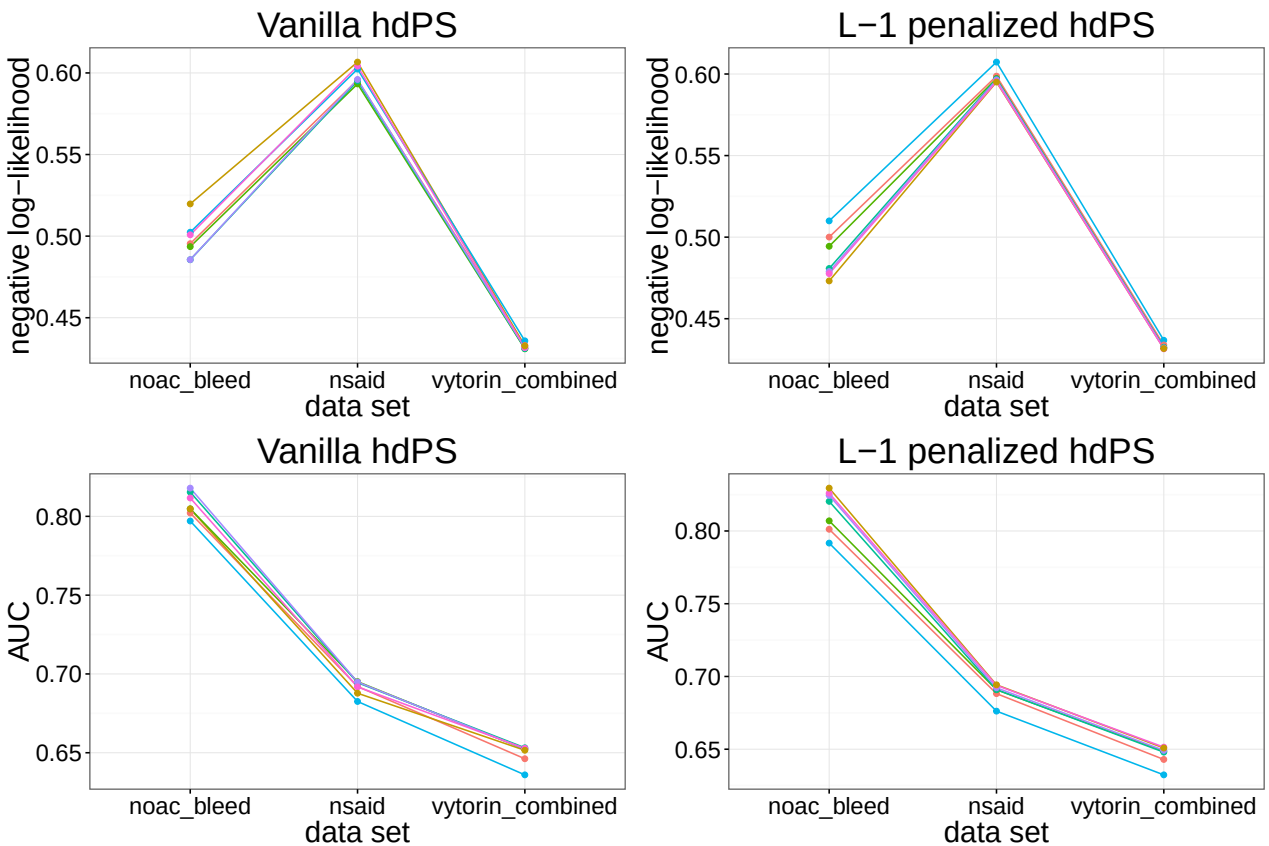


Figure: Unregularized hdPS Compared to Regularized hdPS. The performance is very close for hdPS with/without penalty, which suggests the regularization might not be necessary (on these data sets).

Contribution of the study

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- The primary contribution of this work is that this is the first paper to consider and introduce the novel strategies of combining the SL with the hdPS.
- The other contribution is that this is the most thorough evaluation of ML algorithms and the SL with healthcare claims data.
- Combining the hdPS with SL may be promising for prediction modeling in large healthcare databases, with rich claims data.

Future work

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Variable selection for propensity score estimation with Collaborative Targeted Maximum Likelihood Estimation (C-TMLE).
- Propensity score model selection among a continumm of estimators with C-TMLE.
- Super Learner for ensemble of Convolutional Neural Networks to overcome over-confidence.

Selected References

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- M. van der Laan, and S. Rose. "Targeted learning: causal inference for observational and experimental data." Springer Science & Business Media (2011).
- S. Schneeweiss, J. A. Rassen, R. J. Glynn, J. Avorn , H. Mogun, and M. A. Brookhart. "High-dimensional propensity score adjustment in studies of treatment effects using health care claims data." *Epidemiology* 20.4 (2009): 512.
- M. van der Laan, E. Polley, and A. Hubbard. "Super learner." *Statistical Applications in Genetics and Molecular Biology* 6.1 (2007).
- S. E. Sinisi, E. C. Polley, M. L. Petersen, S. Y. Rhee, and M. J. van der Laan, Super learning: an application to the prediction of HIV-1 drug resistance. *Statistical applications in genetics and molecular biology*, 6(1) (2007).
- S. Dudoit and M. van der Laan. "Asymptotics of cross-validated risk estimation in estimator selection and performance assessment." *Statistical Methodology* 2.2 (2005): 131-154.
- M. van der Laan and S. Dudoit. "Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples." *Bepress* (2003).

Oracle Inequalities

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Define CV selector:

$$\hat{K} = \arg\min_k \mathbb{E}_{B_n}\left(\int (L(O, \hat{\Psi}_k(P_{n,T(v)}))dP_{n,V(v)})\right)$$

- Define oracle selector:

$$\tilde{K} = \arg\min_k \mathbb{E}_{B_n}\left(\int (L(O, \hat{\Psi}_k(P_{n,T(v)}))dP)\right)$$

where B_n is the binary split indicator, P_n is the empirical distribution on the whole learning data, and $P_{n,T(v)}, P_{n,V(v)}$ are the empirical distribution for the training/validation data.

Oracle Inequalities

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

- Let $d_0(\psi, \psi_0) = \mathbb{E}_0(L(O, \psi) - L(O, \psi_0))$ is the risk difference between the candidate estimate and true parameter value ψ_0
- With assumptions:
  1. A1: $L(O, \psi)$ must be uniformly bounded:

$$\sup_{O,\psi} |L(\psi) - L(\psi_0)|(O) \leq M_1 < \infty$$

  2. A2: variance of the ψ_0-centered loss function can be bounded by its expectation uniformly in ψ.

$$\sup_{\psi} \frac{Var_{P_0}(L(O, \psi) - L(O, \psi_0))}{\mathbb{E}_0(L(O, \psi) - L(O, \psi_0))} \leq M_2 < \infty$$

Oracle Inequalities

Super Learner
Cheng Ju
Motivation
High-dimensional Propensity Score Methods
High-dimensional Propensity Score Methods
Review of Super Learner
Data Analysis
Data Description
Results
Discussion
Reference
Appendix

$$\mathbb{E}(d_0(\hat{\Psi}_{\hat{K}(P_n)}, \psi_0)) \leq (1+2\lambda)\,\mathbb{E}(d_0(\hat{\Psi}_{\tilde{K}(P_n)}, \psi_0)) + 2C(\lambda)\frac{1+\log(K(n))}{np} \quad (1)$$

where p is the proportion of the observations in the validation sample, and $C(\lambda) = 2(1+\lambda)^2(M_1/3 + M_2/3)$ is a constant only depend on M_1, M_2, λ (van der Laan et al. 2006).