Targeted Learning with Big Data

Mark van der Laan
UC Berkeley

Center for Philosophy and History of Science
Revisiting the Foundations of Statistics in the Era of Big Data:
Scaling Up to Meet the Challenge

February 20, 2014

Outline

1. Targeted Learning
2. Two stage methodology: Super Learning+ TMLE
3. Definition of Estimation Problem for Causal Effects of Multiple Time Point Interventions
4. Variable importance analysis examples of Targeted Learning
5. Scaling up Targeted Learning to handle Big Data
6. Concluding remarks

Outline

Foundations of the statistical estimation problem

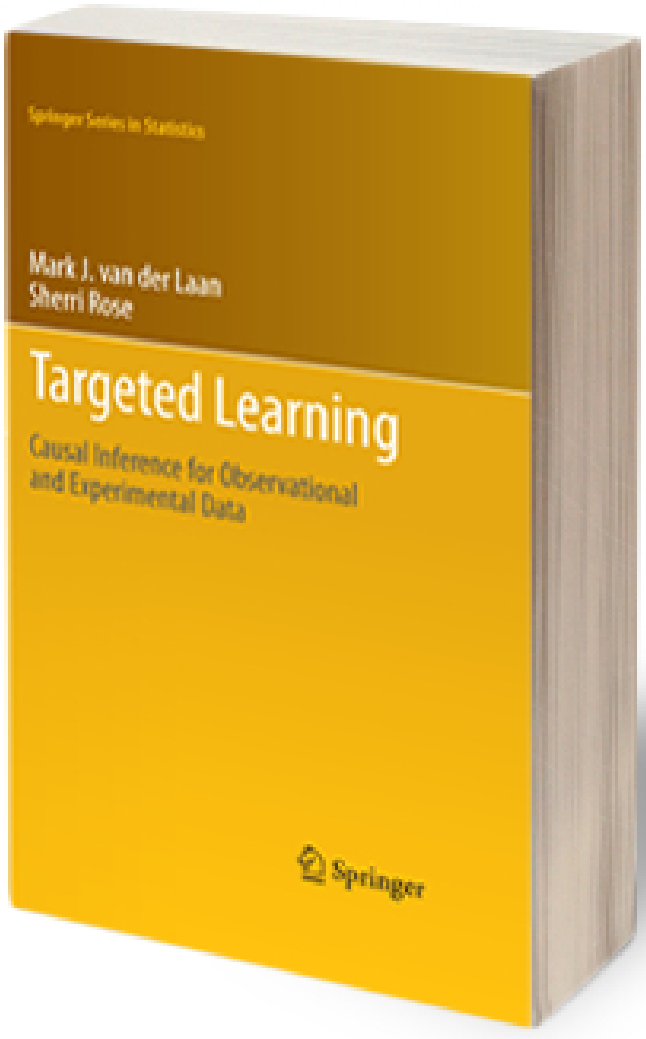
- **Observed data**: Realizations of random variables with a probability distribution.
- **Statistical model**: Set of possible distributions for the data-generating distribution, defined by actual knowledge about the data. e.g. in an RCT, we know the probability of each subject receiving treatment.
- **Statistical target parameter**: Function of the data-generating distribution that we wish to learn from the data.
- **Estimator**: An a priori-specified algorithm that takes the observed data and returns an estimate of the target parameter. Benchmarked by a dissimilarity-measure (e.g., MSE) w.r.t target parameter.
- **Inference:** Establish limit distribution and corresponding statistical inference.

Causal inference

- Non-testable assumptions in addition to the assumptions defining the statistical model. (e.g. the "no unmeasured confounders" assumption).
- Defines causal quantity and establishes identifiability under these assumptions.
- This process generates interesting statistical target parameters.
- Allows for causal interpretation of statistical parameter/estimand.
- Even if we don't believe the non-testable causal assumptions, the statistical estimation problem is still the same, and estimands still have valid statistical interpretations.

Targeted learning

- Define valid (and thus LARGE) statistical semi parametric models and interesting target parameters.
- Exactly deals with statistical challenges of high dimensional and large data sets (Big Data).
- Avoid reliance on human art and nonrealistic (e.g., parametric) models
- Plug-in estimator based on targeted fit of the (relevant part of) data-generating distribution to the parameter of interest
- Semiparametric efficient and robust
- Statistical inference
- Has been applied to: static or dynamic treatments, direct and indirect effects, parameters of MSMs, variable importance analysis in genomics, longitudinal/repeated measures data with time-dependent confounding, censoring/missingness, case-control studies, RCTs, networks.



Targeted Learning Book

Springer Series in Statistics

van der laan & Rose

targetedlearningbook.com

- First Chapter by R.J.C.M. Starmans "Models, Inference, and Truth" provides historical philosophical perspective on Targeted Learning.
- Discusses the erosion of the notion of model and truth throughout history and the resulting lack of unified approach in statistics.
- It stresses the importance of a reconciliation between machine learning and statistical inference, as provided by Targeted Learning.

Outline

Two stage methodology

- Super learning (SL) van der Laan et al. (2007),Polley et al. (2012),Polley and van der Laan (2012)
  - Uses a library of candidate estimators (e.g. multiple parametric models, machine learning algorithms like neural networks, RandomForest, etc.)
  - Builds data-adaptive weighted combination of estimators using cross validation
- Targeted maximum likelihood estimation (TMLE) van der Laan and Rubin (2006)
  - Updates initial estimate, often a Super Learner, to remove bias for the parameter of interest
  - Calculates final parameter from updated fit of the data-generating distribution

Super learning

- No need to chose a priori a particular parametric model or machine learning algorithm for a particular problem
- Allows one to combine many data-adaptive estimators into one improved estimator.
- Grounded by oracle results for loss-function based cross-validation (Van Der Laan and Dudoit (2003),van der Vaart et al. (2006)). Loss function needs to be bounded.
- Performs asymptotically as well as best (oracle) weighted combination, or achieves parametric rate of convergence.

Super learning

Figure: Relative Cross-Validated Mean Squared Error (compared to main terms least squares regression)

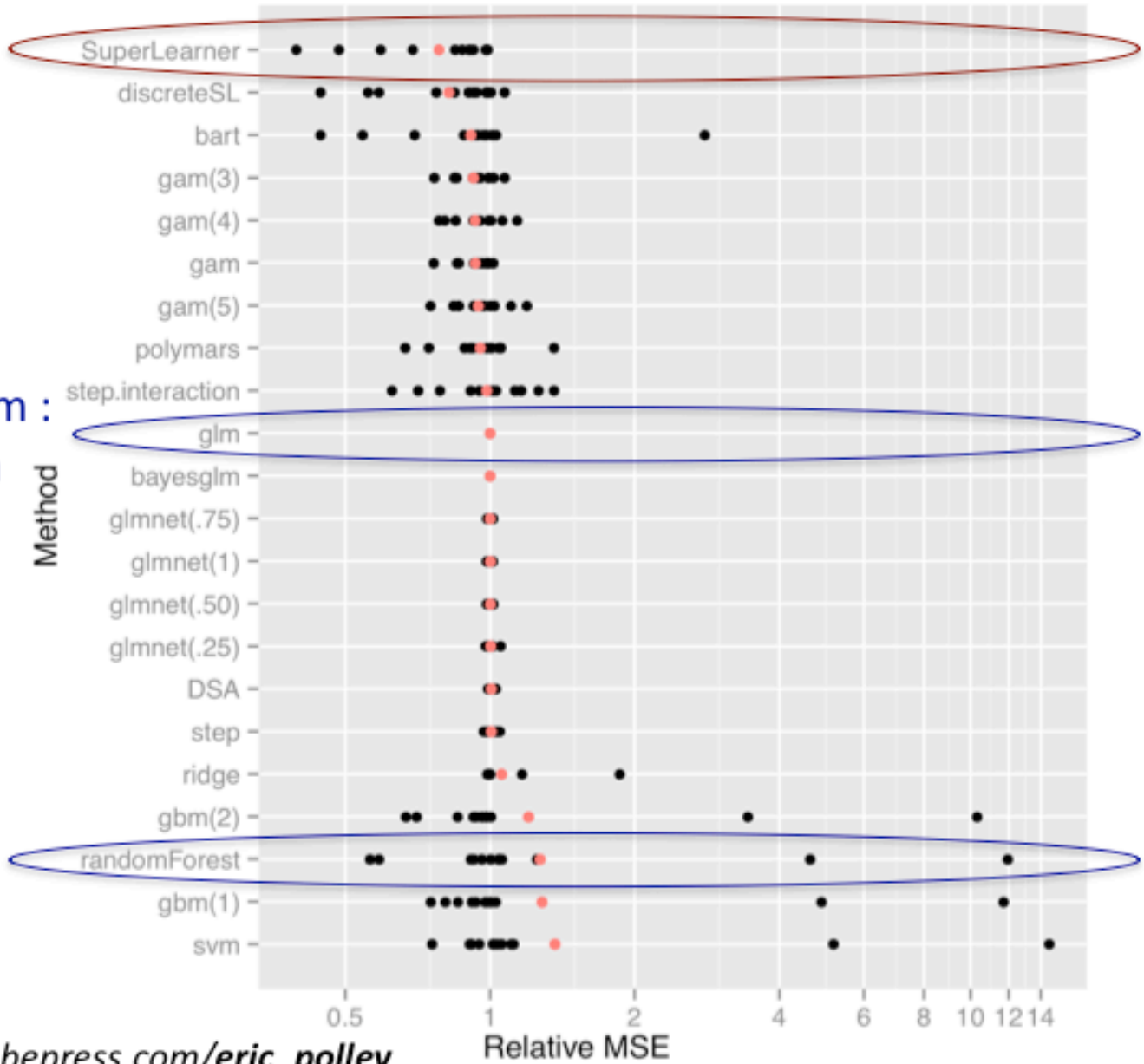
Method	Study 1	Study 2	Study 3	Study 4	Overall
Least Squares	1.00	1.00	1.00	1.00	1.00
LARS	0.91	0.95	1.00	0.91	0.95
D/S/A	0.22	0.95	1.04	0.43	0.71
Ridge	0.96	0.9	1.02	0.98	1.00
Random Forest	0.39	0.72	1.18	0.71	0.91
MARS	0.02	0.82	0.17	0.61	0.38
Super Learner	0.02	0.67	0.16	0.22	0.19

Super learning

Super Learner-
Best weighted combination of algorithms for a given prediction problem

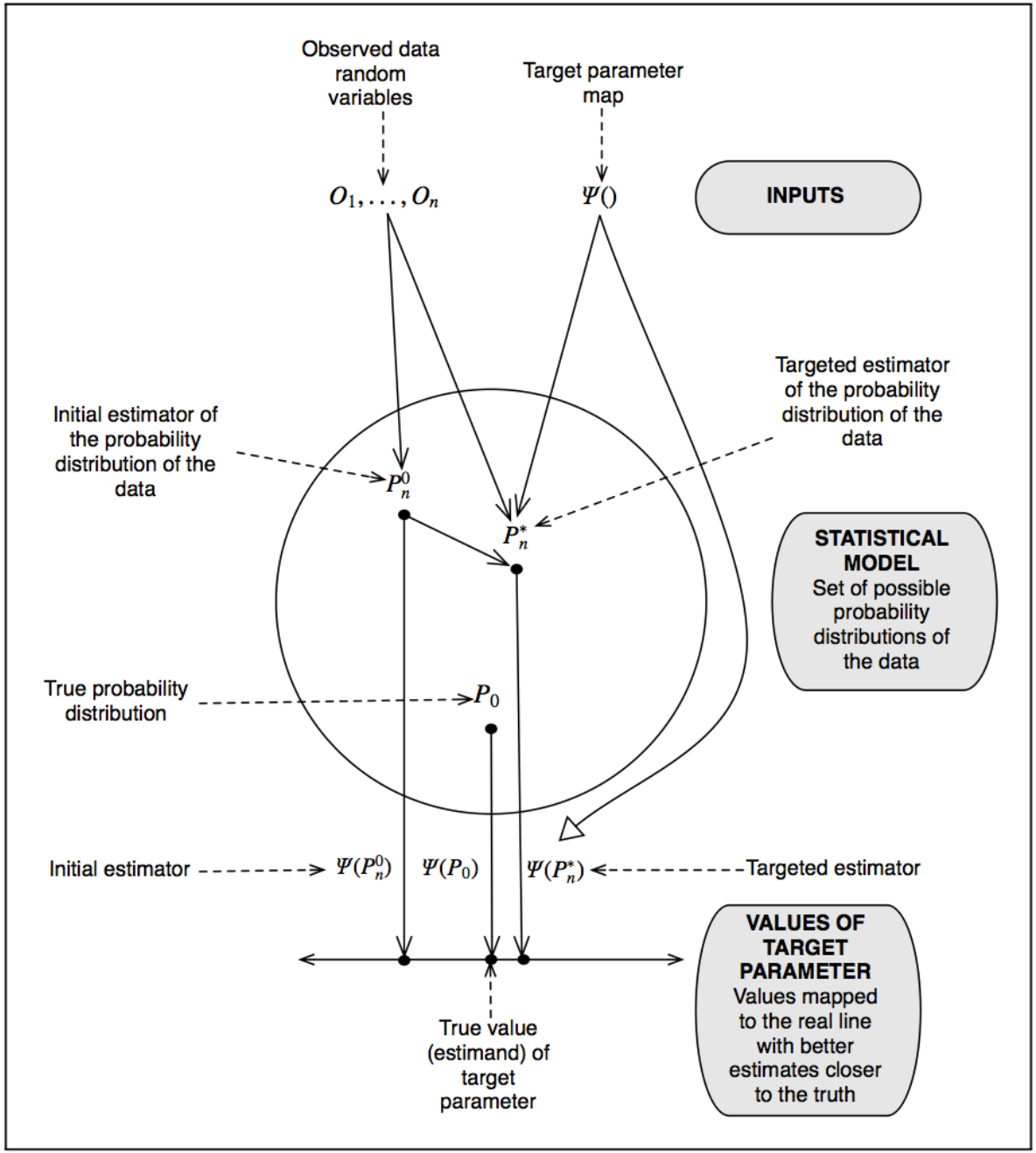
Example algorithm : Linear Main Term Regression

Example algorithm: Random Forest



Technical Report: *works.bepress.com/**eric_polley***

TMLE algorithm



TMLE algorithm: Formal Template

$\Psi(Q_0)$ target parameter

$Q_0 = \arg\min_Q P_0 L(Q) \equiv \int L(Q)(o) dP_0(o)$

$\hat{Q}(P_n)$: Initial estimator, Loss-based SL

$\{\hat{Q}_g(\epsilon) : \epsilon\}$ fluct. model for fitting ψ_0

$\hat{g} = \hat{g}(P_n)$ loss based SL of treatment/cens mech

$\frac{d}{d\epsilon} L(\hat{Q}_{\hat{g}}(\epsilon))\Big|_{\epsilon=0} = D^*(\hat{Q}, \hat{g})$

$\epsilon_n = \arg\min_\epsilon P_n L(\hat{Q}_{\hat{g}}(\epsilon))$

Iterate till convergence: $\hat{Q}^*$

Solves efficient influence curve equation:

$P_n D^*(\hat{Q}^*, \hat{g}) = 0$

TMLE: $\Psi(\hat{Q}^*)$

Outline

General Longitudinal Data Structure

We observe n i.i.d. copies of a longitudinal data structure

$$O = (L(0), A(0), \ldots, L(K), A(K), Y = L(K+1)),$$

where $A(t)$ denotes a discrete valued intervention node, $L(t)$ is an intermediate covariate realized after $A(t-1)$ and before $A(t)$, $t = 0, \ldots, K$, and Y is a final outcome of interest.

For example, $A(t) = (A_1(t), A_2(t))$ could be a vector of two binary indicators of censoring and treatment, respectively.

Likelihood and Statistical Model

The probability distribution P_0 of O can be factorized according to the time-ordering as

$$\begin{aligned} P_0(O) &= \prod_{t=0}^{K+1} P_0(L(t) \mid Pa(L(t))) \prod_{t=0}^{K} P_0(A(t) \mid Pa(A(t))) \\ &\equiv \prod_{t=0}^{K+1} Q_{0,L(t)}(O) \prod_{t=0}^{K} g_{0,A(t)}(O) \\ &\equiv Q_0 g_0, \end{aligned}$$

where $Pa(L(t)) \equiv (\bar{L}(t-1), \bar{A}(t-1))$ and $Pa(A(t)) \equiv (\bar{L}(t), \bar{A}(t-1))$ denote the parents of $L(t)$ and $A(t)$ in the time-ordered sequence, respectively. The g_0-factor represents the intervention mechanism: e..g, treatment and right-censoring mechanism.

Statistical Model: We make no assumptions on Q_0, but could make assumptions on g_0

Statistical Target Parameter: G-computation Formula for Post-Intervention Distribution

- Let

$$P^d(l) = \prod_{t=0}^{K+1} Q^d_{L(t)}(\bar{l}(t)), \tag{1}$$

  where $Q^d_{L(t)}(\bar{l}(t)) = Q_{L(t)}(l(t) \mid \bar{l}(t-1), \bar{A}(t-1) = \bar{d}(t-1))$.
- Let $L^d = (L(0), L^d(1), \ldots, Y^d = L^d(K+1))$ denote the random variable with probability distribution P^d.
- This is the so called G-computation formula for the post-intervention distribution corresponding with the dynamic intervention d.

Example: When to switch a failing drug regimen in HIV-infected patients

- **Observed data on unit**

$$O = (L(0), A(0), L(1), A(1), \ldots, L(K), A(K), A_2(K)Y),$$

where $L(0)$ is baseline history, $A(t) = (A_1(t), A_2(t))$, $A_1(t)$ is indicator of switching drug regimen, $A_2(t)$ is indicator of being right-censored, $t = 0, \ldots, K$, and Y is indicator of observing death by time $K + 1$.

- Define interventions nodes $A(0), \ldots, A(K)$ and interventions dynamic rules d_θ that switch when CD4-count drops below θ, and enforces no-censoring.
- Our target parameter is defined as projection of $(E(Y^{d_\theta}(t)) : t, d)$ onto a working model $m_\beta(\theta)$ with parameter β.

A Sequential Regression G-computation Formula (Bang, Robins, 2005)

- By the iterative conditional expectation rule (tower rule), we have

$$E_{P^d} Y^d = E \ldots E(E(Y^d \mid \bar{L}^d(K)) \mid L^d(K-1)) \ldots \mid L(0)).$$

- In addition, the conditional expectation, given $\bar{L}^d(K)$ is equivalent with conditioning on $\bar{L}(K), \bar{A}(K-1) = \bar{d}(K-1)$.

- In this manner, one can represent $E_{P^d} Y^d$ as an iterative conditional expectation, first take conditional expectation, given $\bar{L}^d(K)$ (equivalent with $\bar{L}(K), \bar{A}(K-1)$), then take the conditional expectation, given $\bar{L}^d(K-1)$ (equivalent with $\bar{L}(K-1), \bar{A}(K-2)$), and so on, until the conditional expectation given $L(0)$, and finally take the mean over $L(0)$.
- We developed a targeted plug-in estimator/TMLE of general summary measures of "dose-response" curves $(EY_d : d \in \mathcal{D})$ (Petersen et al., 2013, van der Laan, Gruber 2012).

Outline

Variable Importance: Problem Description (Diaz, Hubbard)

- Around 800 patients that entered the emergency room with severe trauma
- About 80 physiological and clinical variables were measured at 0, 6, 12, 24, 48, and 72 hours after admission
- Objective is predicting the most likely medical outcome of a patient (e.g., survival), and provide an ordered list of the covariates that drive this prediction (variable importance).
- This will help doctors decide what variables are relevant at each time point.
- Variables are subject to missingness
- Variables are continuous, variable importance parameter is

$$\Psi(P_0) \equiv E_0\{E_0(Y \mid A + \delta, W) - E_0(Y \mid A, W)\}$$

for user-given value δ.

Variable Importance: Results

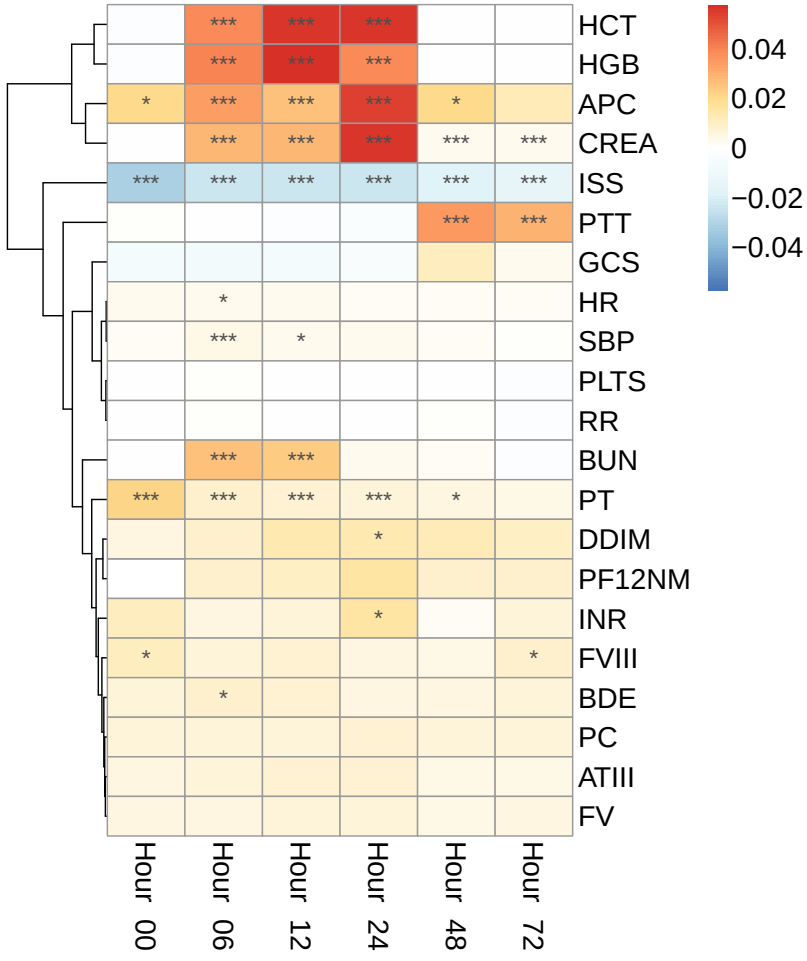


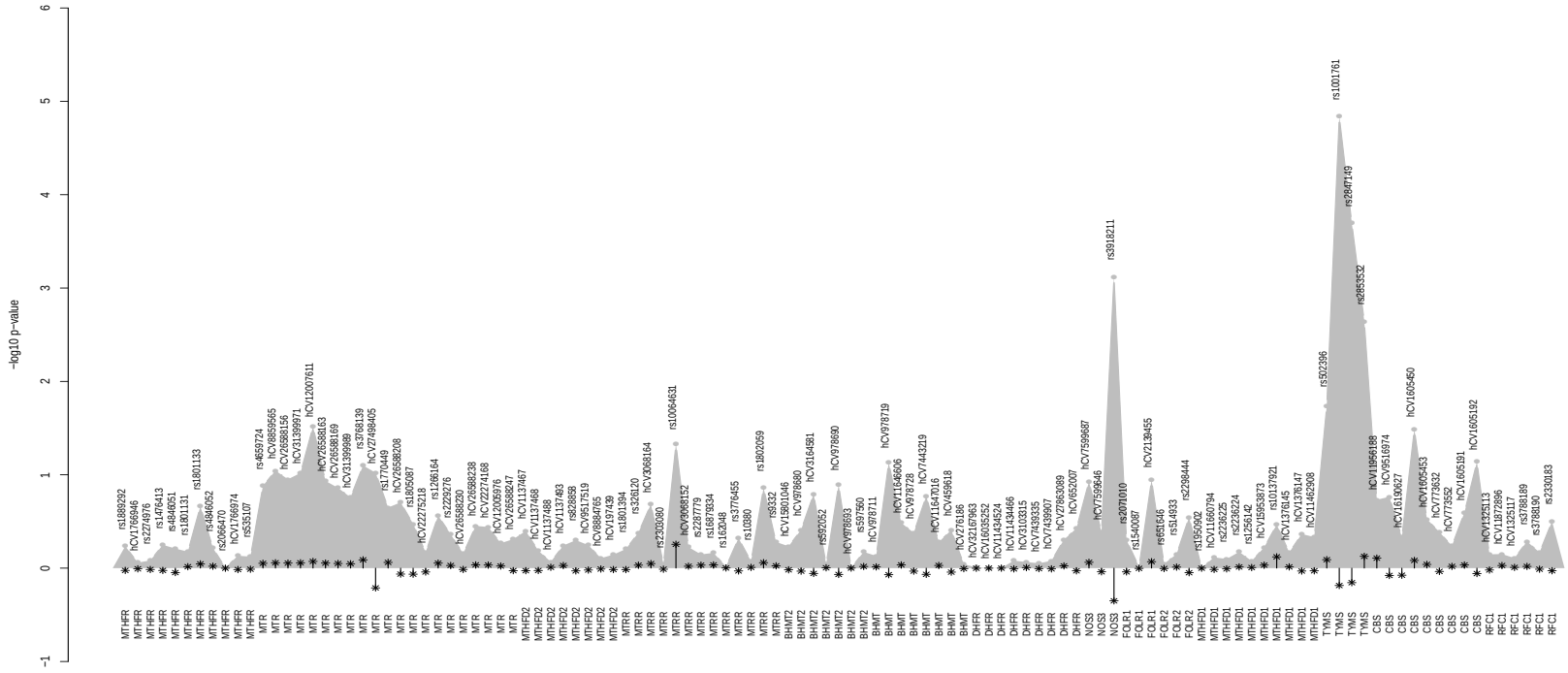
Figure: Effect sizes and significance codes

TMLE with Genomic Data

- 570 case-control samples on spina bifida
- We want to identify associated genes to spina bifida from 115 SNPs.
- In the original paper Shaw et. al. 2009, a univariate analysis was performed.

- The original analysis missed rs1001761 and rs2853532 in TYMS gene because they are closely linked with counteracting effects on spina bifida.
- With TMLE, signals from these two SNPs were recovered.
- In TMEL, Q0 was obtained from LASSO, and $g(W)$ is obtained from a simple regression of SNP on its two flanking markers to account for confounding effects of neighborhood markers.

TMLE p-values for SNPs



Outline

Targeted Learning of Data Dependent Target Parameters (vdL, Hubbard, 2013)

- Define algorithms that map data into a target parameter: $\Psi_{P_n} : \mathcal{M} \rightarrow \mathbb{R}^d$, thereby generalizing the notion of target parameters.
- Develop methods to obtain statistical inference for $\Psi_{P_n}(P_0)$ or $1/V \sum_{v=1}^{V} \Psi_{P_{n,v}}(P_0)$, where $P_{n,v}$ is the empirical distribution of parameter-generating sample corresponding with v-th sample split. We have developed cross-validated TMLE for the latter data dependent target parameter, without any additional conditions.
- In particular, this generalized framework allows us to generate a subset of target parameters among a massive set of candidate target parameters, while only having to deal with multiple testing for the data adaptively selected set of target parameters.
- Thus, much more powerful than regular multiple testing for a fixed set of null hypotheses.

Online Targeted MLE: Ongoing work

- Order data if not ordered naturally.
- Partition in subsets numbered from 1 to K.
- Initiate initial estimator and TMLE based on first subset.
- Update initial estimator based on second subset, and update TMLE based on second subset.
- Iterate till last subset.
- Final estimator is average of all stage specific TMLEs.
- In this manner, for each subset number of calculations is bounded by number of observations in subset and total computation time increases linearly in number of subsets.
- One can still prove asymptotic efficiency of this online TMLE.

Outline

Concluding remarks

- Sound foundations of statistics are in place (Data is random variable, Model, Target Parameter, Inference based on Limit Distribution), but these have eroded over many decades.
- However, MLE had to be revamped into TMLE to deal with large models.
- Big Data asks for development of fast TMLE without giving up on statistical properties: e.g., Online TMLE.
- Big Data asks for research teams that consists of top statisticians and computer scientists, beyond subject matter experts.
- Philosophical soundness of proposed methods are hugely important and should become a norm.

References I

H. Bang and J. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

M. Petersen, J. Schwab, S. Gruber, N. Blaser, M. Schomaker, and M. van der Laan. Targeted minimum loss based estimation of marginal structural working models. *Journal of Causal Inference*, submitted, 2013.

E. Polley and M. van der Laan. *SuperLearner: Super Learner Prediction*, 2012. URL `http://CRAN.R-project.org/package=SuperLearner`. R package version 2.0-6.

E. Polley, S. Rose, and M. van der Laan. Super learning. *Chapter 3 in Targeted Learning, van der Laan,Rose (2012)*, 2012.

References II

M. Schnitzer, E. Moodie, M. van der Laan, R. Platt, and M. Klein. Marginal structural modeling of a survival outcome with targeted maximum likelihood estimation. *Biometrics*, to appear, 2013.

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. *UC Berkeley Division of Biostatistics Working Paper Series*, page 130, 2003.

M. J. van der Laan and S. Gruber. Targeted Minimum Loss Based Estimation of Causal Effects of Multiple Time Point Interventions. *The International Journal of Biostatistics*, 8(1), Jan. 2012. ISSN 1557-4679. doi: $10.1515/1557\text{-}4679.1370$.

References III

M. J. van der Laan and D. Rubin. Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, 2(1), Jan. 2006. ISSN 1557-4679. doi: $10.2202/1557\text{-}4679.1043$.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), Jan. 2007. ISSN 1544-6115. doi: $10.2202/1544\text{-}6115.1309$.

A. van der Vaart, S. Dudoit, and M. van der Laan. Oracle inequalities for multi-fold cross-validation. *Stat Decis*, 24(3): 351–371, 2006.