Assessing Variable Importance Nonparametrically using Machine Learning Techniques

Brian D. Williamson,
Peter Gilbert, Noah Simon, Marco Carone

Student Paper Competition, WNAR 2017

26 June 2017

Variable importance

- Data $O_1, O_2, \ldots, O_n$ from unknown distribution $P_0 \in \mathcal{M}$
    - $O_i := (X_i, Y_i)$
    - Covariate vector $X_i := (X_{i1}, X_{i2}, \ldots, X_{ip}) \in \mathbb{R}^p$
    - Outcome $Y_i \in \mathbb{R}$
- Estimate $\mu_{P_0}(x) := E_{P_0}(Y \mid X = x)$
- Which features contribute most to variation in $\mu_{P_0}(x)$?
    - Consider $\mu_{P_0,s}(x) := E_{P_0}(Y \mid X_{(-s)} = x_{(-s)})$
    - $X_{(-s)}$ is the vector with the element(s) in $s \subseteq \{1, 2, \ldots, p\}$ removed

Variable importance (continued)

- Fundamental questions:
  - How do we estimate μ_{P_0} and $\mu_{P_0,s}$?
  - How do we quantify variable importance?
- Approaches:
  - Parametric, e.g., ANOVA; must be correctly specified
  - Model-agnostic:
    - Technique-specific measures, e.g., random forests [Breiman (2001)]
    - Technique-agnostic measures [Doksum and Samarov (1995)], [van der Laan (2006), Chambaz et al. (2012), Sapp et al. (2014)]

Our goals

Flexible, Interpretable

- Estimate μ_{P_0} and $\mu_{P_0,s}$ using state-of-the-art methods
- Estimate a scientifically meaningful parameter consistently and efficiently
- Properly quantify the uncertainty in our estimates

The parameter of interest

- Additional proportion of variability in Y explained by including X_s in the regression:

$$\psi_{0,s} \equiv \Psi_s(P_0) := \frac{\int \left\{E_{P_0}(Y \mid X = x) - E_{P_0}(Y \mid X_{(-s)} = x_{(-s)})\right\}^2 dP_0(x)}{var_{P_0}(Y)}$$

- $\Psi_s(P_0)$ is a property of the data generating mechanism
- Interpretation does not change with estimating procedure
- Equivalent to difference in R^2 between the two regressions:

$$\Psi_s(P_0) = \frac{E_{P_0}[\{Y - \mu_{P_0}(X)\}^2]}{var_{P_0(Y)}} - \frac{E_{P_0}[\{Y - \mu_{P_0,s}(X)\}^2]}{var_{P_0}(Y)}$$

Efficient influence function?

- MLE $\hat{\theta}_n$ of θ_0; information $I(\theta_0)$, score $\dot{\ell}(\theta_0 \mid X)$
- Let $\tilde{\ell}(\theta_0 \mid X) = I^{-1}(\theta_0)\dot{\ell}(\theta_0 \mid X)$:
  - This is the efficient influence function (EIF) for θ_0
  - $\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\ell}(\theta_0 \mid X_i) + o_p(1)$
  - $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N\left[0, E_{P_0}\left\{\tilde{\ell}(\theta_0 \mid X)^2\right\}\right] = N\{0, I^{-1}(\theta_0)\}$
- Given an EIF for a nonparametric parameter:
  - Estimator with influence function = EIF is efficient
  - Can use similar distribution theory to parametric case

The EIF for $\Psi_s(P)$ relative to $\mathcal{M}$

$$\begin{aligned}
\mu_P(x) &= E_P(Y \mid X = x) \\
\mu_{P,s}(x) &= E_P(Y \mid X_{(-s)} = x_{(-s)}) \\
\phi_s(P) &= \int \{\mu_P(x) - \mu_{P,s}(x)\}^2 \, dP(x)
\end{aligned}$$

Then

$$\begin{aligned}
o \mapsto D^*_{P,s}(o) := & \frac{2\{y - \mu_P(x)\}\{\mu_P(x) - \mu_{P,s}(x)\} + \{\mu_P(x) - \mu_{P,s}(x)\}^2}{var_P(Y)} \\
& - \phi_s(P) \left\{ \frac{y - E_P(Y)}{var_P(Y)} \right\}^2
\end{aligned}$$

Asymptotic expansion

- Estimate the relevant components of P_0 using $\widehat{P}_n$
- Linearize Ψ using the EIF $D^*_{P,s}$ and use the empirical $\mathbb{P}_n$:

$$
\begin{aligned}
\Psi_s(\widehat{P}_n) - \Psi_s(P_0) = & \int D^*_{\widehat{P}_n,s}(o)d(\widehat{P}_n - P_0)(o) + R_s(\widehat{P}_n, P_0) \\
= & \frac{1}{n}\sum_{i=1}^{n} D^*_{P_0,s}(O_i) \\
& + \int \{D^*_{\widehat{P}_n,s}(o) - D^*_{P_0,s}(o)\} d(\mathbb{P}_n - P_0)(o) \\
& + R_s(\widehat{P}_n, P_0) - \frac{1}{n}\sum_{i=1}^{n} D^*_{\widehat{P}_n,s}(O_i)
\end{aligned}
$$

■ linear term;	(1^{st} order)
■ empirical process term;	(2^{nd} order)
■ remainder term;	(2^{nd} order)
■ problem term!	(irregular)

A naive estimator of $\Psi_s(P_0)$

$$\psi_{0,s} \equiv \Psi_s(P_0) = \frac{\int \left\{\mu_{P_0}(x) - \mu_{P_0,s}(x)\right\}^2 dP_0(x)}{var_{P_0}(Y)}$$

- Given estimators $\hat{\mu}(x)$ and $\hat{\mu}_s(x)$
- Plug in:

$$\hat{\psi}_{\text{naive},s} = \frac{n^{-1}\sum_{i=1}^{n}\left\{\hat{\mu}(X_i) - \hat{\mu}_s(X_i)\right\}^2}{n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}$$

Problems with the naive estimator

$$\Psi_s(\widehat{P}_n) - \Psi_s(P_0) = \frac{1}{n}\sum_{i=1}^{n} D^*_{P_0,s}(O_i) + R_s(\widehat{P}_n, P_0) - \frac{1}{n}\sum_{i=1}^{n} D^*_{\widehat{P}_n,s}(O_i) + \int \{D^*_{\widehat{P}_n,s}(o) - D^*_{P_0,s}(o)\} d(\mathbb{P}_n - P_0)(o)$$

- "Bias" incurred from estimating components of P_0
- Generally neither efficient nor regular and asymptotically linear

The one-step estimator

- Remove bias and get regularity, asymptotic linearity, and efficiency by adding on $\frac{1}{n}\sum_{i=1}^{n} D^*_{\widehat{P}_{n,s}}(O_i)$:

$$\hat{\psi}_{n,s} = \hat{\psi}_{\text{naive, s}} + \frac{1}{n}\sum_{i=1}^{n} D^*_{\widehat{P}_{n,s}}(O_i),$$

or equivalently

$$\hat{\psi}_{n,s} = \hat{\psi}_{\text{naive, s}} + \frac{n^{-1}\sum_{i=1}^{n} 2\{Y_i - \hat{\mu}(X_i)\}\{\hat{\mu}(X_i) - \hat{\mu}_s(X_i)\}}{n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}$$

Asymptotic behavior of the one-step estimator

Under some regularity conditions,

$$\sqrt{n}(\hat{\psi}_{n,s} - \psi_{0,s}) = n^{-1/2} \sum_{i=1}^{n} D^*_{P_0,s}(O_i) + o_P(1)$$

and

$$\sqrt{n}(\hat{\psi}_{n,s} - \psi_{0,s}) \rightarrow_d N\left[0, E_{P_0}\left\{D^*_{P_0,s}(O)^2\right\}\right].$$

- Consistent, regular, efficient
- Regularity conditions:
  - $\psi_{0,s} \neq 0$
  - $\hat{\mu}$, $\hat{\mu}_s$ converge quickly enough
  - $D^*_{\hat{P}_n,s}$ falls in a P_0-Donsker class with probability tending to one
- Estimate variance of $\hat{\psi}_{n,s}$ empirically

Simulations with a low-dimensional vector of covariates

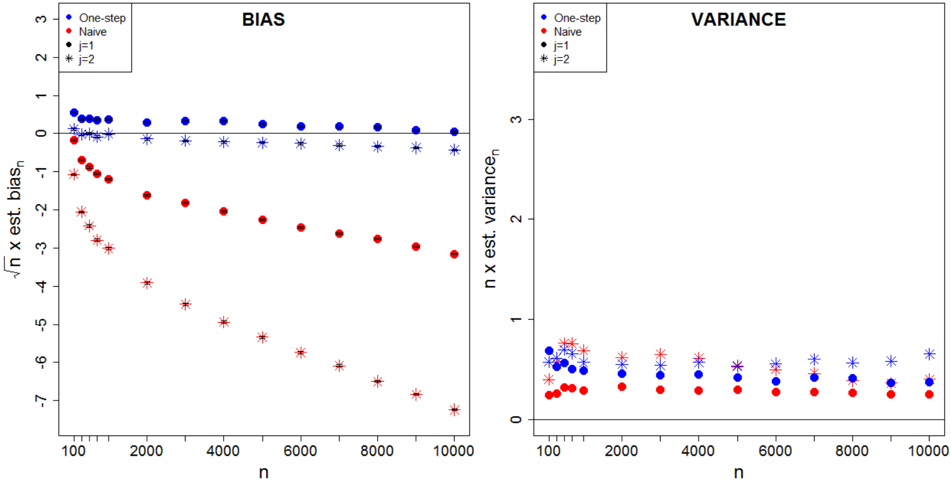
- Data:

$$X_1, X_2 \overset{iid}{\sim} Unif(-1, 1) \text{ and } \epsilon \sim N(0, 1) \text{ independent of } (X_1, X_2)$$

$$Y = X_1^2 \left(X_1 + \frac{7}{5} \right) + \frac{25}{9} X_2^2 + \epsilon$$

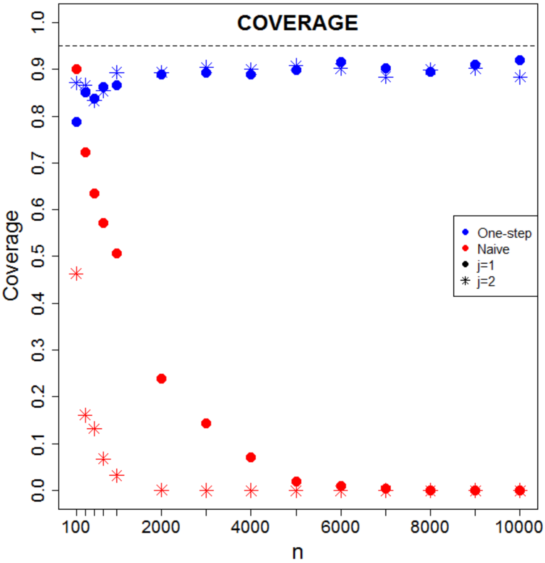
- Truths: $\psi_{0,1} \approx 0.158$, $\psi_{0,2} \approx 0.342$
- Locally-constant loess, five-fold CV to obtain optimal bandwidths
- Percentile bootstrap for naive confidence intervals

Results



BIAS
One-step
Naive
j=1
j=2
3
2
1
0
-1
-2
-3
-4
-5
-6
-7
$\sqrt{n}$ x est. bias$_n$
100
2000
4000
6000
8000
10000
n
VARIANCE
One-step
Naive
j=1
j=2
3
2
1
0
n x est. variance$_n$
100
2000
4000
6000
8000
10000
n

Results



COVERAGE
Coverage
n
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
100
2000
4000
6000
8000
10000
One-step
Naive
j=1
j=2

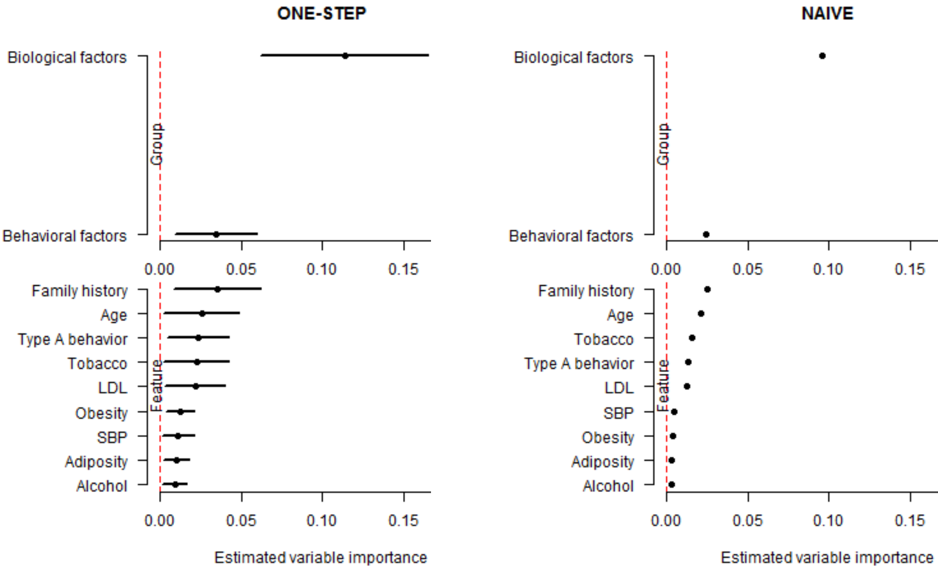
The CORIS data [Rousseaw et al. (1983)]

$n = 462$, outcome = presence of MI

- Behavioral:
  - tobacco consumption,
  - alcohol consumption,
  - type-A behavior
- Biological:
  - systolic blood pressure,
  - LDL cholesterol,
  - adiposity,
  - obesity,
  - family history,
  - age

Super learner [van der Laan et al. (2007)] with boosted trees, elastic net, GAMs, random forests, and five-fold CV

Results from the CORIS data



Conclusions

- Interpretable: Additional proportion of variability in Y explained by including X_s in the estimation technique
- Flexible: Valid CIs with state-of-the-art methods!
- Consistently and efficiently estimate a property of the data generating mechanism
- Reasonable performance in simulation and in data analysis
- Implemented in R package `vimp` and Python package `vimpy`
- Future work:
  - dealing with a boundary null hypothesis,
  - working in a structured model (e.g., additive models),
  - nested case-control study data,
  - censoring

References

[1] Breiman, L. Random forests. *Machine Learning*, 2001.

[2] Chambaz A, Neuvial P, and van der Laan MJ. Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 2012.

[3] Doksum K and Samarov A. Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *The Annals of Statistics*, 1995.

[4] Rousseauw J, Du Plessis J, Benade A, Jordann P, Kotze J, Jooste P, and Ferreira J. Coronary risk factor screening in three rural communities. *South African Medical Journal*, 1983.

[5] Sapp S, van der Laan MJ, and Page K. Targeted estimation of binary variable importance measures with interval-censored outcomes. *The International Journal of Biostatistics*, 2014.

[6] van der Laan MJ. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2006.

[7] van der Laan MJ, Polley EC, and Hubbard AE. Super Learner. *UC Berkeley Division of Biostatistics Working Paper Series*, 2007.