Nonparametric Variable Importance Assessment using Machine Learning Techniques

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

JSM 2018

30 July 2018

Motivation

What is the importance of different biological measurements for predicting the presence or absence of myocardial infarction (MI)?

$X_1 =$ heart rate

$X_2 =$ blood sugar

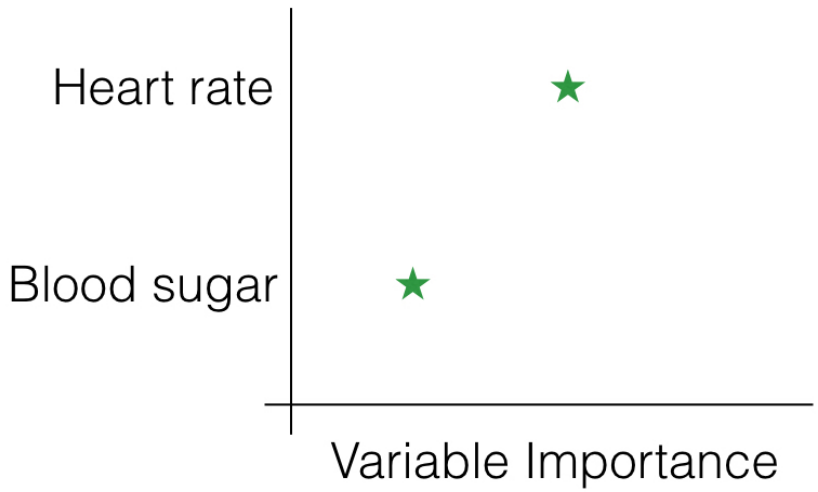
$Y =$ presence or absence of MI

Motivation

What is the importance of different biological measurements for predicting presence or absence of MI?

We need:

- A definition of variable importance with minimal assumptions

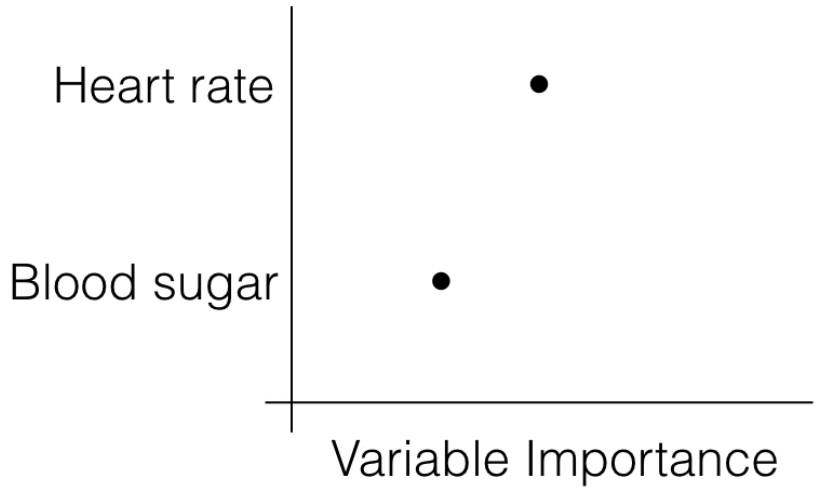


Motivation

What is the importance of different biological measurements for predicting presence or absence of MI?

We need:

- A definition of variable importance with minimal assumptions
- A method that:
  - Estimates variable importance

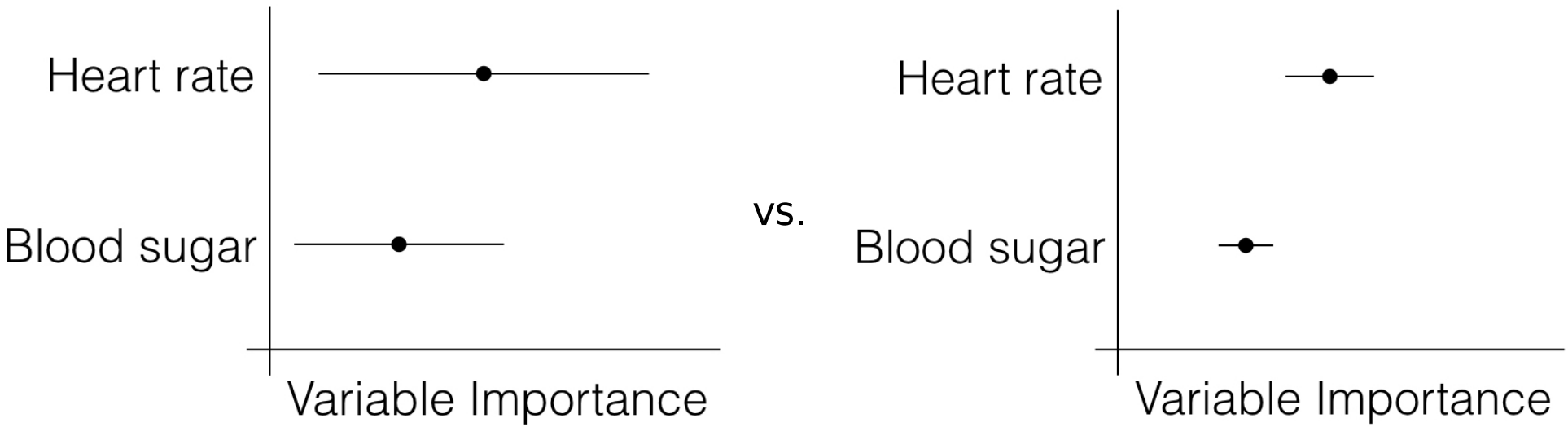


Motivation

What is the importance of different biological measurements for predicting presence or absence of MI?

We need:

- A definition of variable importance with minimal assumptions
- A method that:
  - Estimates variable importance
  - Provides valid uncertainty assessment for our estimates



Motivation

What is the importance of different biological measurements for predicting presence or absence of MI?

We need:

- A definition of variable importance with minimal assumptions
- A method that:
  - Estimates variable importance
  - Provides valid uncertainty assessment for our estimates
  - May be used with flexible estimation procedures

Variable importance

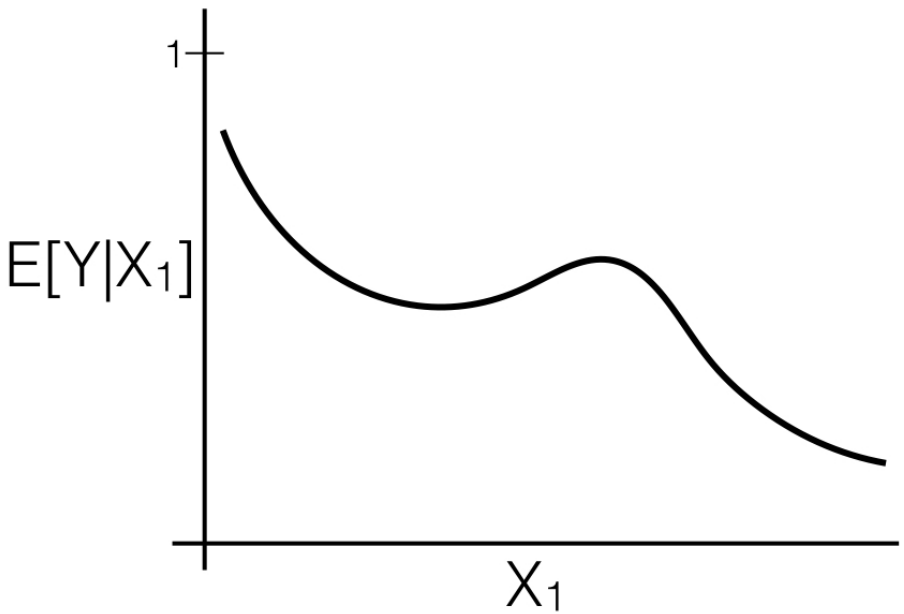
- Data $O_1, O_2, \ldots, O_n$ from unknown distribution $P_0 \in \mathcal{M}$
  - $O_i := (X_i, Y_i)$
  - Covariate vector $X_i := (X_{i1}, X_{i2}, \ldots, X_{ip}) \in \mathbb{R}^p$
  - Outcome $Y_i \in \mathbb{R}$
- Estimate $\mu_{P_0}(x) := E_{P_0}(Y \mid X = x)$
- Which features contribute most to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

Existing variable importance methods

Method	Nonparametric estimation method	Uncertainty assessment	Fully general estimation
Parametric, e.g., ANOVA	—	✓	—
Technique-specific measures, e.g., random forests	✓	—	✓
Nonparametric variable importance using existing estimation procedures	✓	✓	—

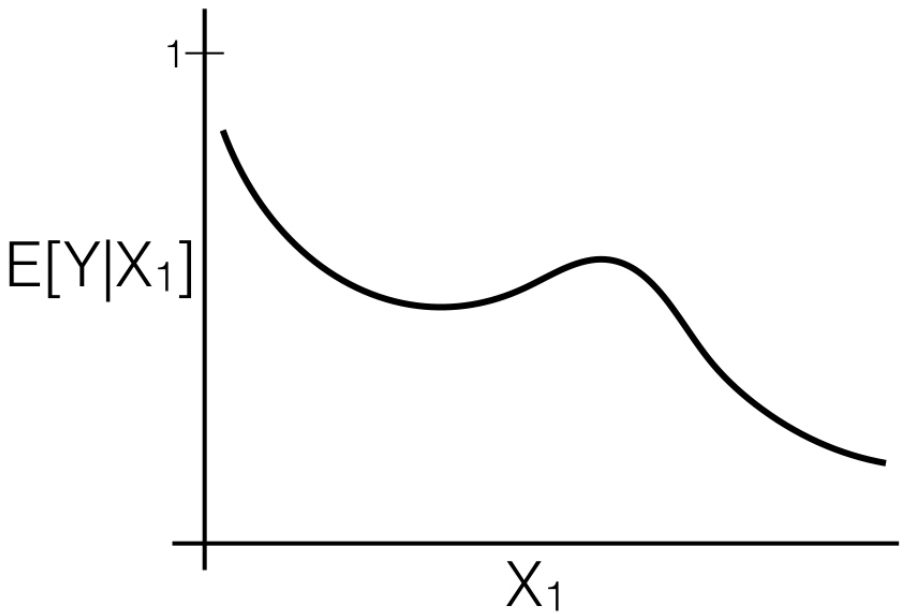
Variable importance: example

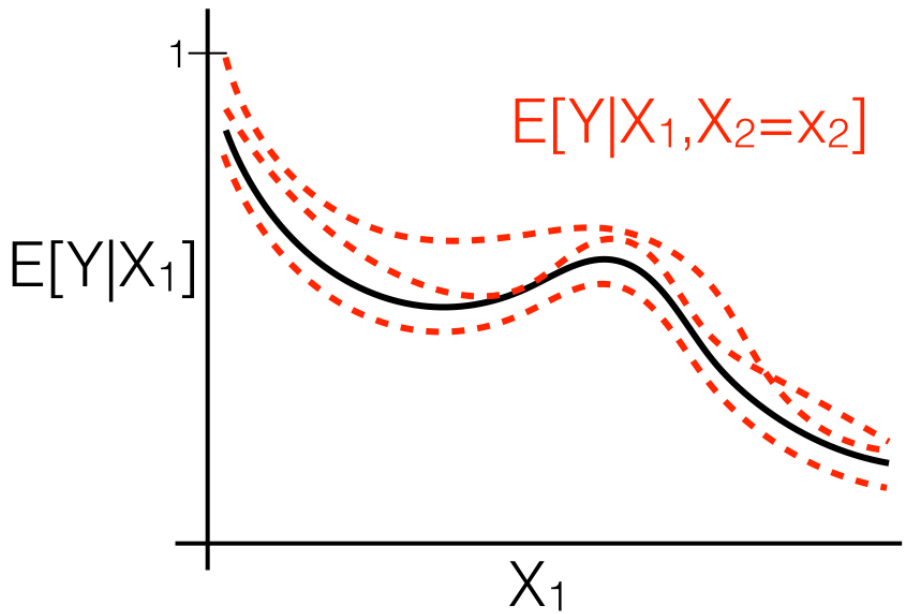
What is the importance of $X_1 =$ heart rate vs. $X_2 =$ blood sugar for predicting $Y =$ presence/absence of MI?



Variable importance: example

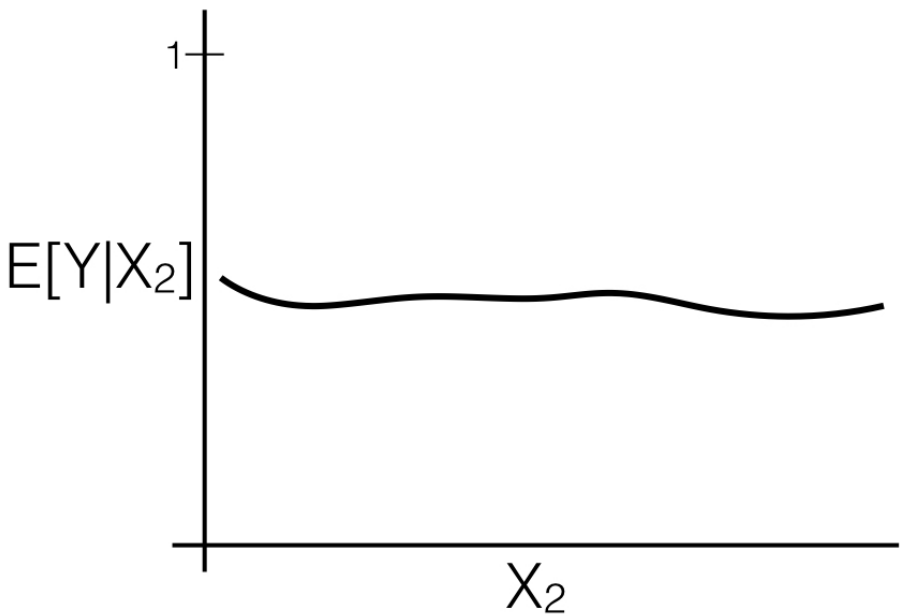
What is the importance of $X_1 =$ heart rate vs. $X_2 =$ blood sugar for predicting $Y =$ presence/absence of MI?





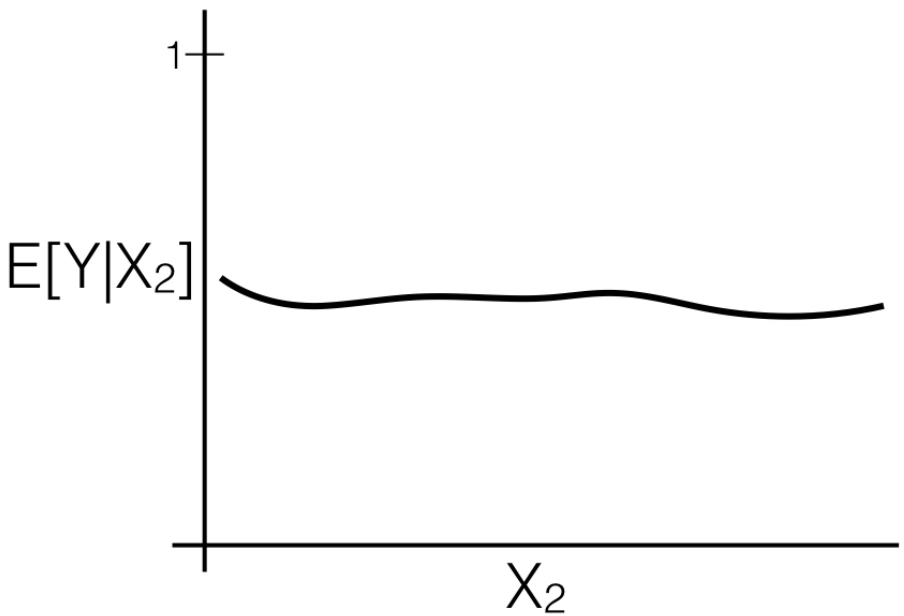
Variable importance: example

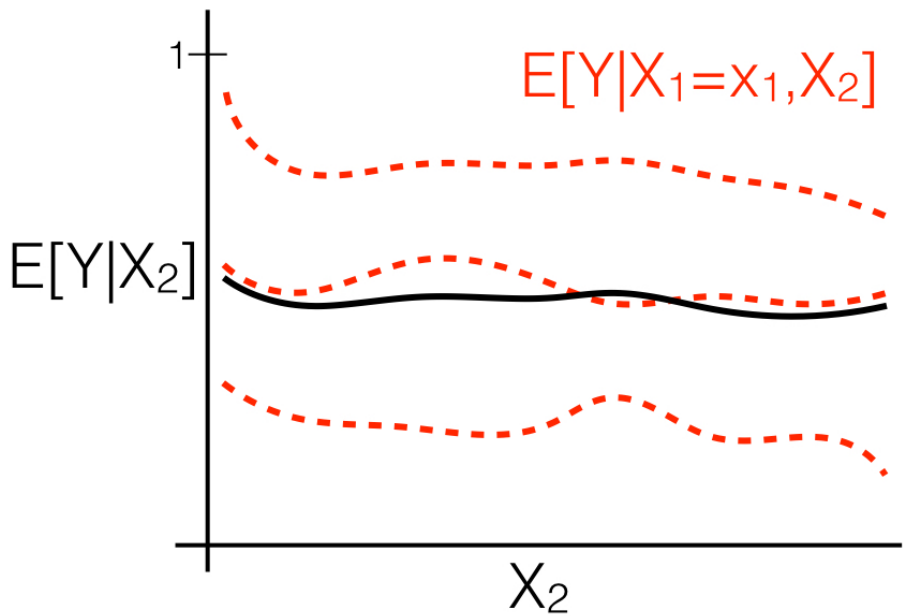
What is the importance of $X_1 =$ heart rate vs. $X_2 =$ blood sugar for predicting $Y =$ presence/absence of MI?



Variable importance: example

What is the importance of $X_1 =$ heart rate vs. $X_2 =$ blood sugar for predicting $Y =$ presence/absence of MI?





Our contribution

We propose a procedure that:

- estimates a scientifically meaningful parameter consistently and efficiently, while
- estimating μ_{P_0} and $\mu_{P_0,s}$ using state-of-the-art methods, and
- properly quantifies the uncertainty in our estimates

The parameter of interest

The importance of X_s relative to $X_{(-s)}$ for predicting Y:

$$\psi_{0,s} := \frac{E_{P_0}\left[\{\mu_{P_0}(X) - \mu_{P_0,s}(X)\}^2\right]}{\mathrm{Var}_{P_0}(Y)}.$$

Interpretation:

- additional proportion of variability in Y explained by including X_s in the regression
- does not change with estimating procedure
- Equivalent to difference in R^2 between the two regressions:

$$\frac{E_{P_0}[\{Y - \mu_{P_0}(X)\}^2]}{\mathrm{Var}_{P_0}(Y)} - \frac{E_{P_0}[\{Y - \mu_{P_0,s}(X)\}^2]}{\mathrm{Var}_{P_0}(Y)}$$

$\psi_{0,s}$ is a **property of $\mathbf{P_0}$**, not any particular algorithm.

Statistical inference

Using a **population-based**, **model-agnostic** variable importance measure allows us to perform **statistical inference**.

We do this by borrowing from classical parametric theory:

- 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
  - $\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\ell}(\theta_0 \mid X_i) + o_p(1)$
  - $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N\left[0, E_{P_0}\left\{\tilde{\ell}(\theta_0 \mid X)^2\right\}\right] = N\{0, I^{-1}(\theta_0)\}$
- 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 $\mathcal{M}$

Define $\Psi_s(P) := \frac{\Phi_s(P)}{\text{Var}_P(Y)}$, where

$$\mu_P(x) = E_P(Y \mid X = x) \text{ and } \mu_{P,s}(x) = E_P(Y \mid X_{(-s)} = x_{(-s)})$$

$$\Phi_s(P) = E_P\left[\{\mu_P(x) - \mu_{P,s}(x)\}^2\right]$$

Then

$$o \mapsto D^*_{P,s}(o) := \frac{2\{y - \mu_P(x)\}\{\mu_P(x) - \mu_{P,s}(x)\} + \{\mu_P(x) - \mu_{P,s}(x)\}^2}{\text{Var}_P(Y)}$$

$$- \Phi_s(P)\left\{\frac{y - E_P(Y)}{\text{Var}_P(Y)}\right\}^2$$

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 $\mathbb{P}_n$:

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

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

A naive estimator of $\psi_{0,s}$

$$\psi_{0,s} = \frac{E_{P_0}\left(\{\mu_{P_0}(X) - \mu_{P_0,s}(X)\}^2\right)}{\mathrm{Var}_{P_0}(Y)}$$

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

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

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

Our proposed corrected estimator

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

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

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 proposed 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

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

- Consistent, regular, efficient
- Regularity conditions:
  - $\psi_{0,s} \neq 0$
  - $\hat{\mu}$, $\hat{\mu}_s$ converge quickly enough to μ_{P_0}, $\mu_{P_0,s}$
  - $D^*_{\hat{P}_n,s}$ 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

For data (Y, X_1, X_2) with distribution

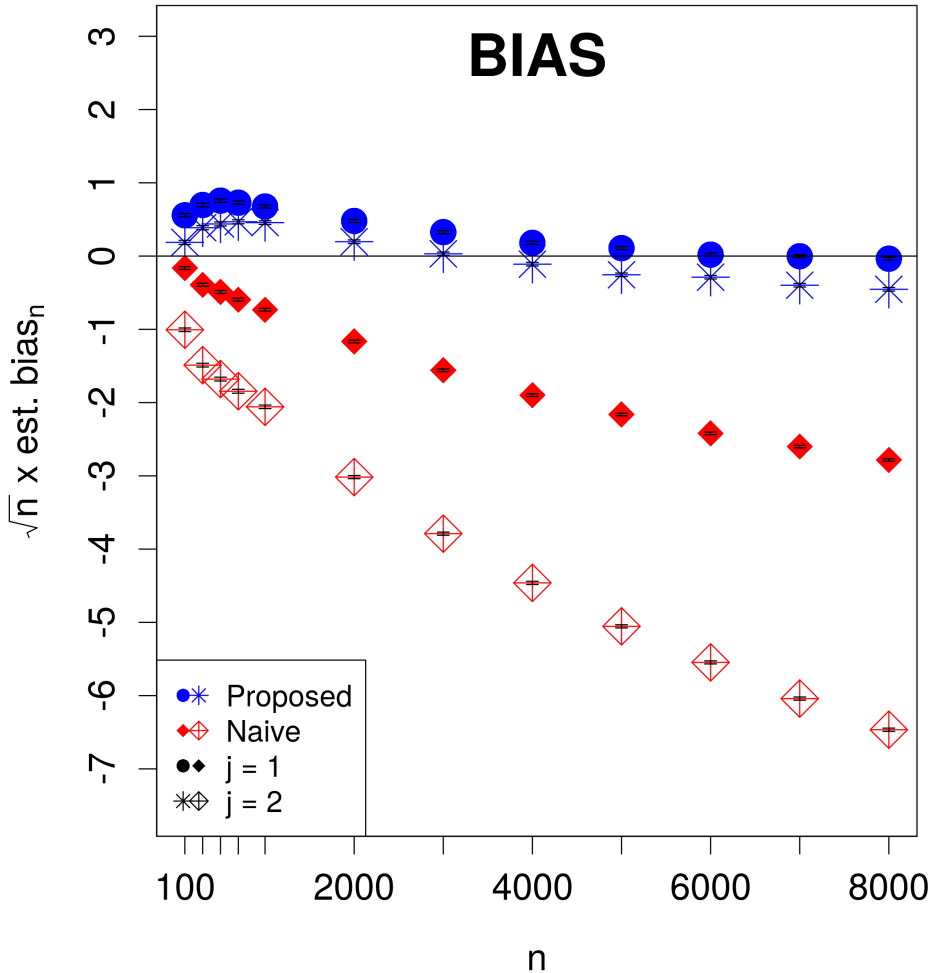
$X_1, X_2 \overset{iid}{\sim} Unif(-1, 1)$ and $\epsilon \sim N(0, 1)$ independent of (X_1, X_2)

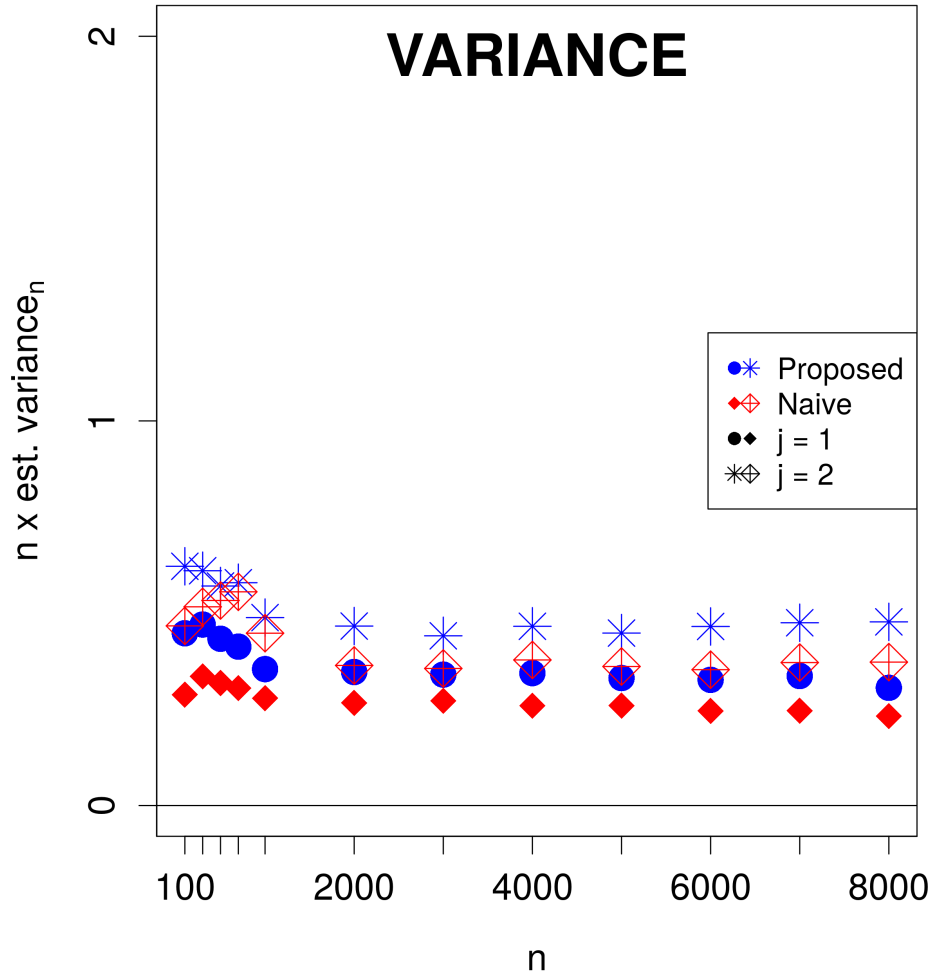
$$Y = X_1^2 \left(X_1 + \frac{7}{5} \right) + \frac{25}{9} X_2^2 + \epsilon$$

estimate the importance of X_1 and X_2.

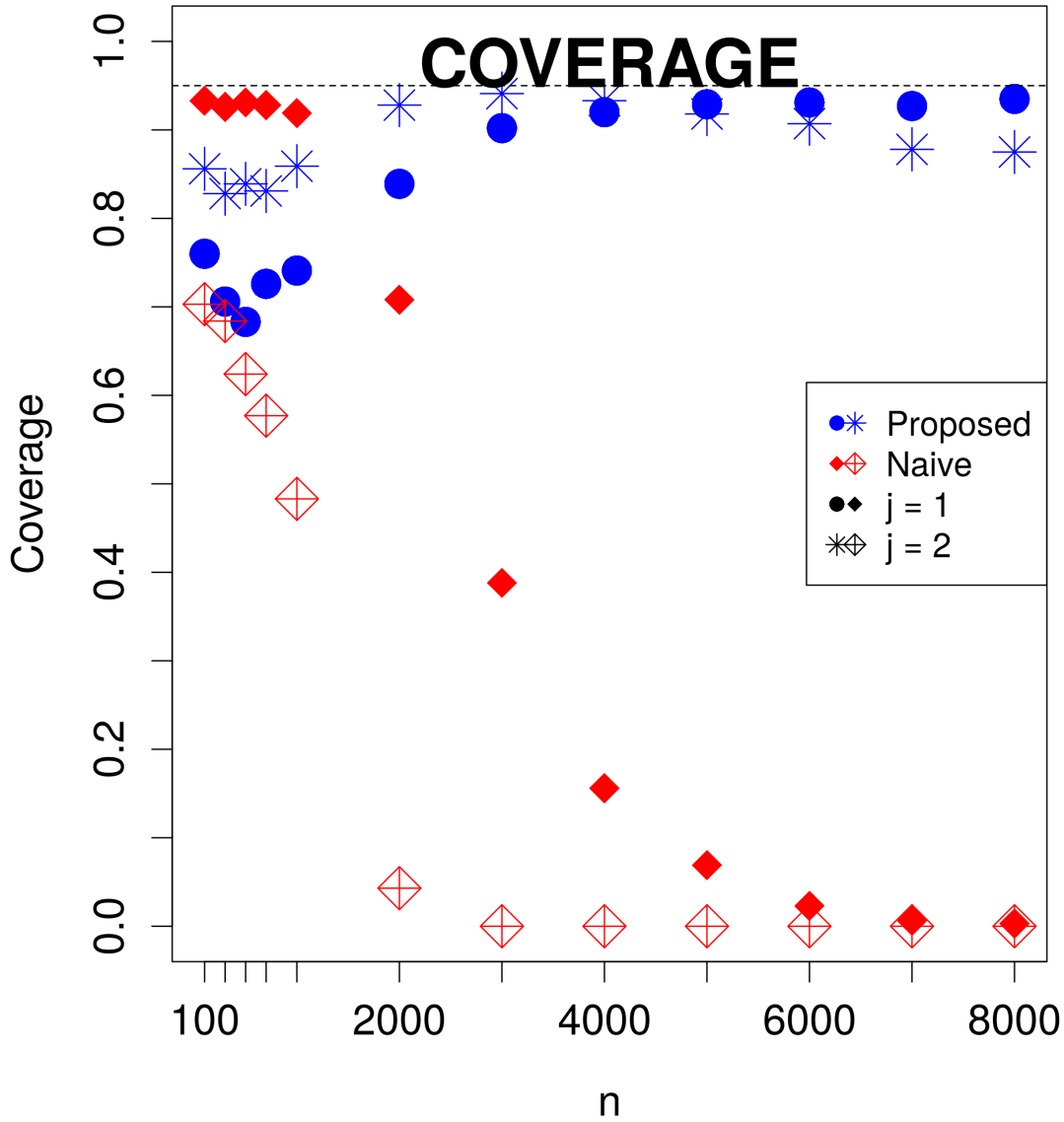
- Truths: $\psi_{0,1} \approx 0.158$, $\psi_{0,2} \approx 0.342$
- Locally-constant loess, five-fold CV for bandwidth
- Percentile bootstrap for naive confidence intervals

Results

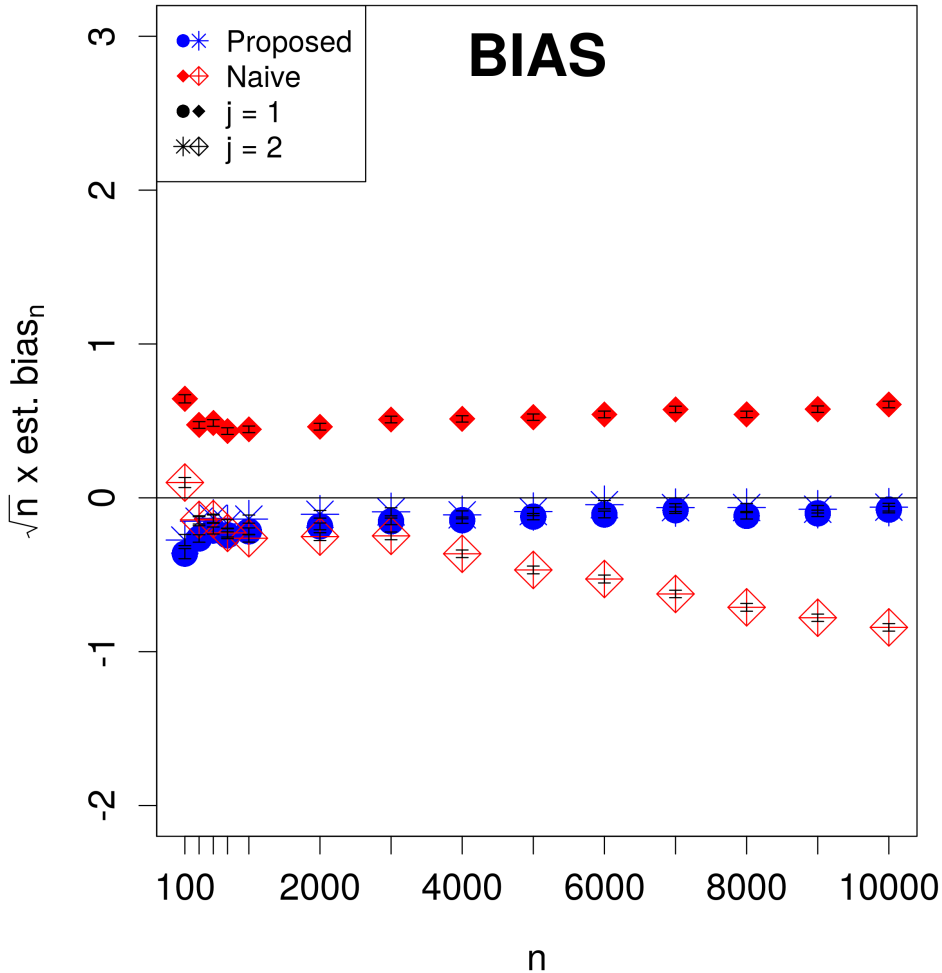

BIAS
$\sqrt{n}$ x est. bias$_n$
3
2
1
0
-1
-2
-3
-4
-5
-6
-7
100
2000
4000
6000
8000
n
Proposed
Naive
j = 1
j = 2

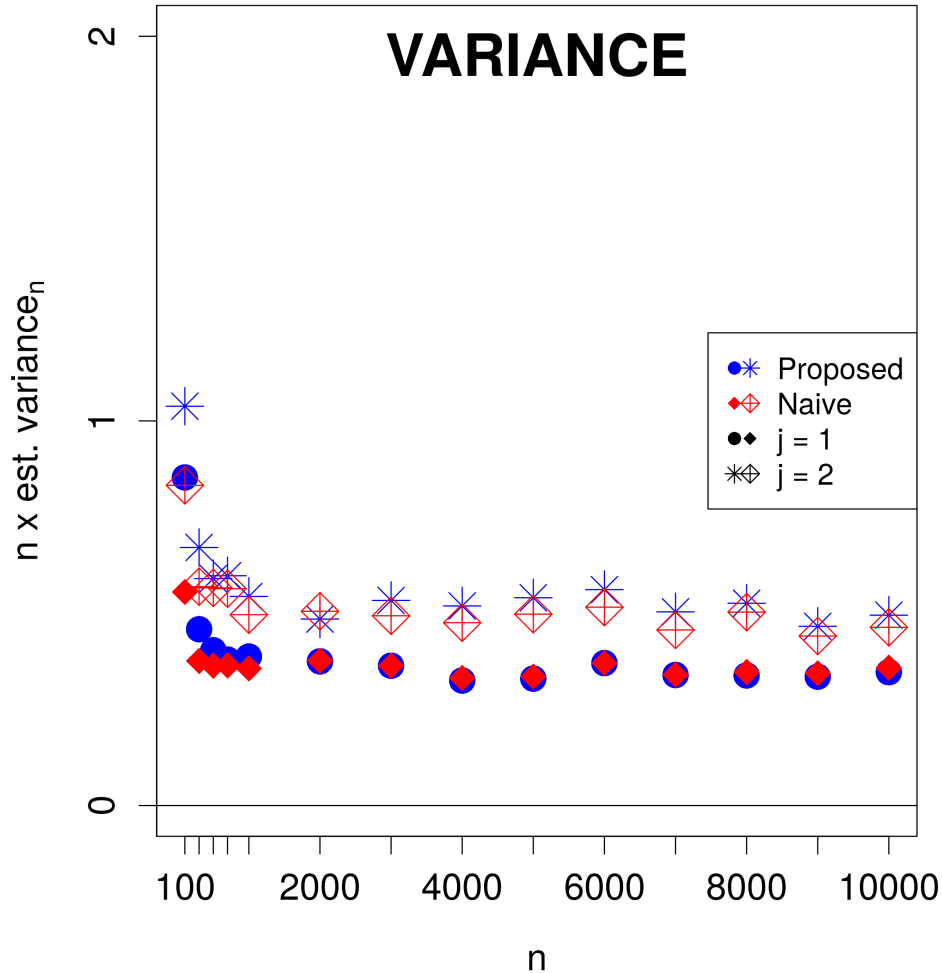

VARIANCE
n x est. variance$_n$
2
1
0
100
2000
4000
6000
8000
n
Proposed
Naive
j = 1
j = 2

Results

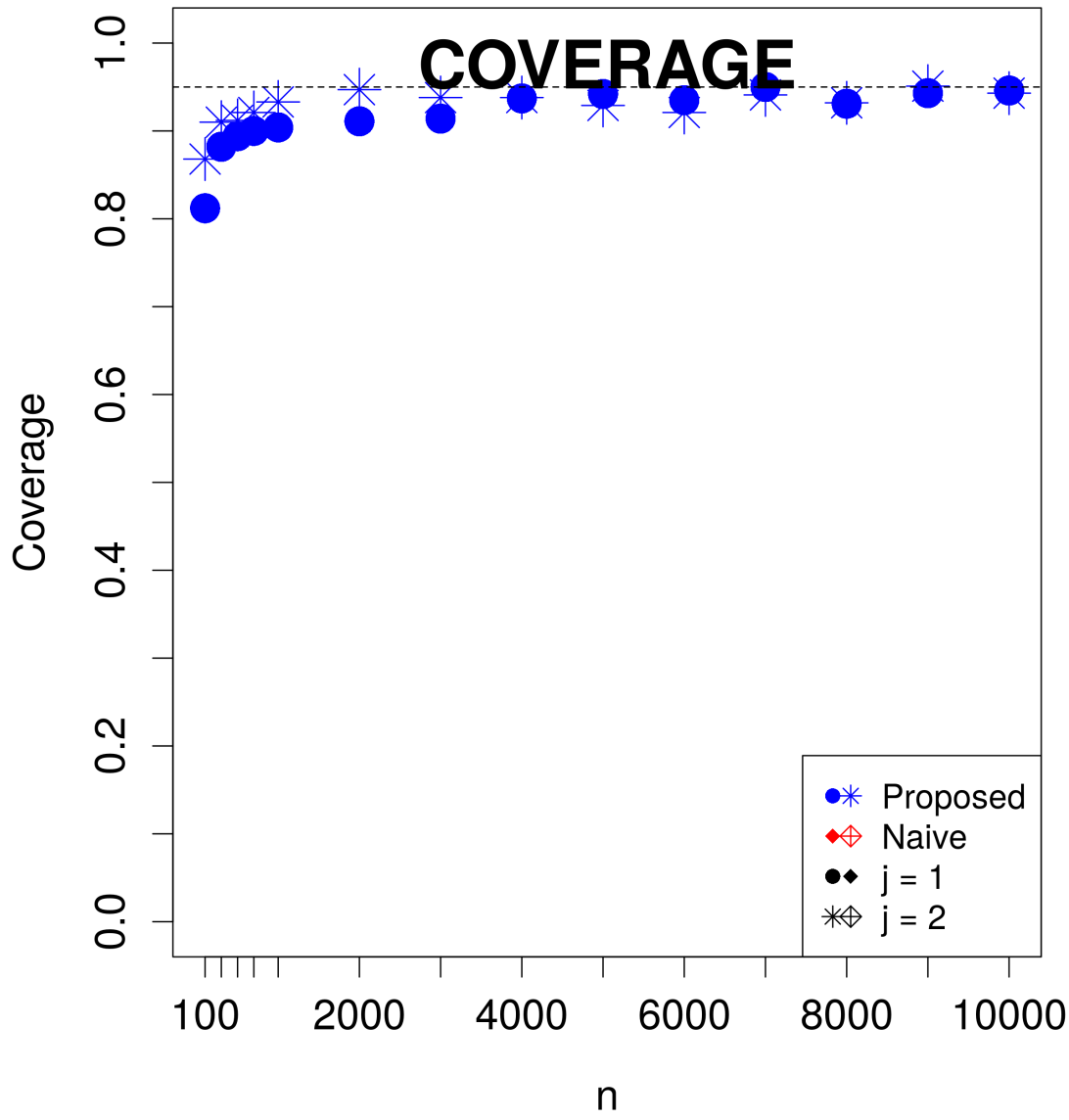

COVERAGE
Coverage
0.0
0.2
0.4
0.6
0.8
1.0
100
2000
4000
6000
8000
n
Proposed
Naive
j = 1
j = 2

Results (cross-validated estimator)





Results (cross-validated estimator)



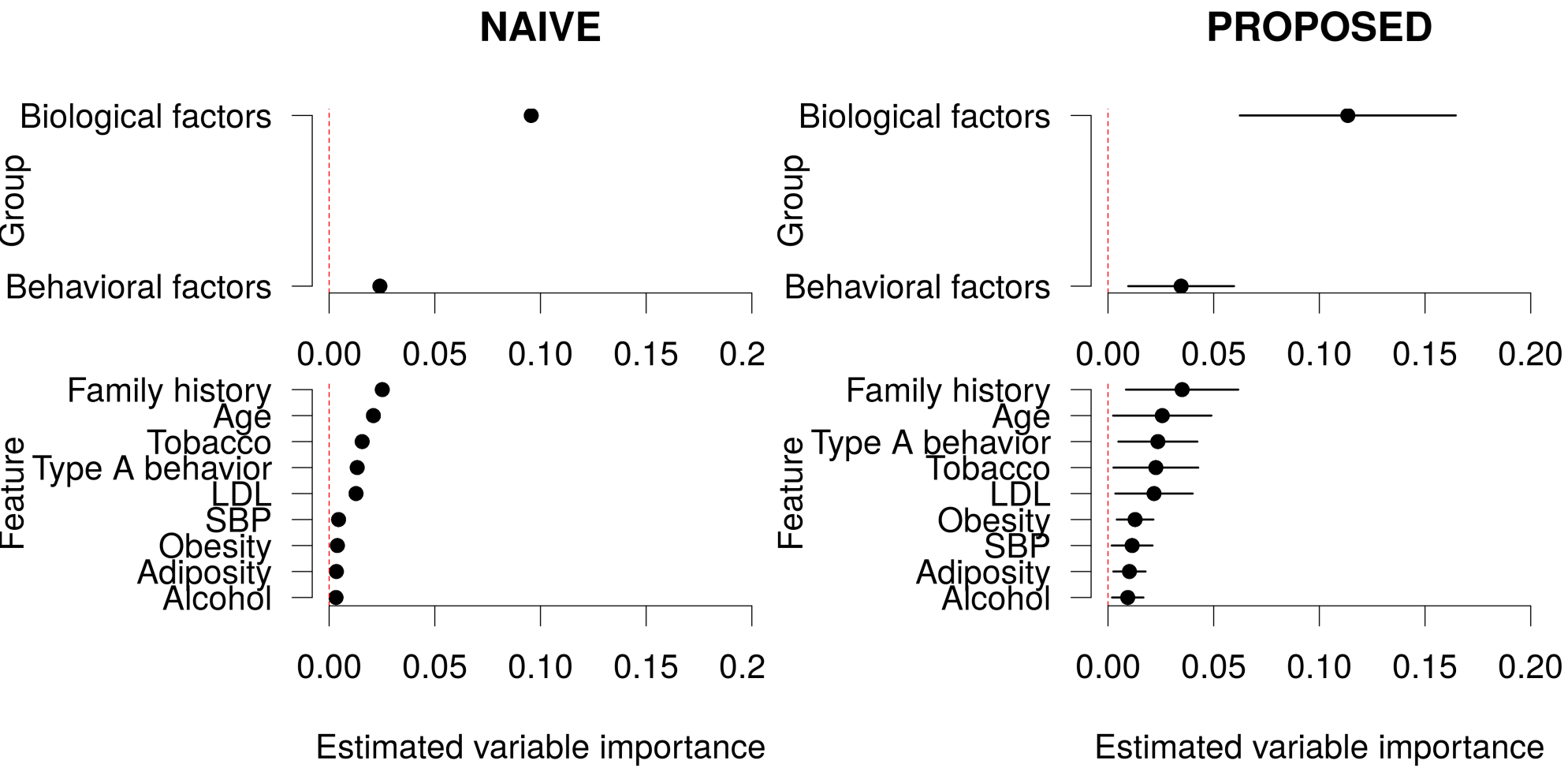
The CORIS data [Rousseaw et al. (1983)]

$n = 462$, 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

We propose a procedure for estimating the **difference in population $\mathbf{R^2}$** when including X_s and removing X_s, where:

- we estimate importance consistently and efficiently,
- obtain valid CIs, and
- estimate the conditional means using state-of-the-art methods

Future work:

- dealing with a boundary null hypothesis,
- working in a structured model (e.g., additive models),
- nested case-control study data,
- censoring

Thank you!

CRAN: package `vimp` **PyPI:** package `vimpy`

https://github.com/bdwilliamson/vimp

Preprint: https://biostats.bepress.com/uwbiostat/paper422/

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] Lei J, G'Sell M, Tibshirani R, and Wasserman L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 2018.

[5] 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.

[6] 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.

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

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