Inference for model-agnostic longitudinal variable importance

Brian D. Williamson, PhD

Kaiser Permanente Washington Health Research Institute

18 December, 2023

https://bdwilliamson.github.io/talks

Acknowledgments

This work was done in collaboration with:

Susan Shortreed

KAISER PERMANENTE®

Kaiser Permanente Washington
Health Research Institute

Erica Moodie

McGill

School of Population and Global Health

Epidemiology, Biostatistics
and Occupational Health

Variable importance: what and why

What is variable importance?

Variable importance: what and why

What is variable importance?

- Quantification of "contributions" of a variable (or a set of variables)

Variable importance: what and why

What is variable importance?

- Quantification of "contributions" of a variable (or a set of variables)

  Traditionally: contribution to predictions

Variable importance: what and why

What is variable importance?

- Quantification of "contributions" of a variable (or a set of variables)

  Traditionally: contribution to predictions

- Useful to distinguish between contributions of predictions...

Variable importance: what and why

What is variable importance?

- Quantification of "contributions" of a variable (or a set of variables)

  Traditionally: contribution to predictions

- Useful to distinguish between contributions of predictions...
  - (extrinsic importance) ... by a given (possibly black-box) algorithm [e.g., Breiman, (2001)]

Variable importance: what and why

What is variable importance?

- Quantification of "contributions" of a variable (or a set of variables)

  Traditionally: contribution to predictions

- Useful to distinguish between contributions of predictions...
  - (extrinsic importance) ... by a given (possibly black-box) algorithm [e.g., Breiman, (2001)]
  - (intrinsic importance) ... by best possible (i.e., oracle) algorithm [e.g., van der Laan (2006)]

Variable importance: what and why

What is variable importance?

- Quantification of "contributions" of a variable (or a set of variables)

  Traditionally: contribution to predictions

- Useful to distinguish between contributions of predictions...
  - (extrinsic importance) ... by a given (possibly black-box) algorithm [e.g., Breiman, (2001)]
  - (intrinsic importance) ... by best possible (i.e., oracle) algorithm [e.g., van der Laan (2006)]
- Our work focuses on interpretable, model-agnostic intrinsic importance

Variable importance: what and why

What is variable importance?

- Quantification of "contributions" of a variable (or a set of variables)

  Traditionally: contribution to predictions

- Useful to distinguish between contributions of predictions...
  - (extrinsic importance) ... by a given (possibly black-box) algorithm [e.g., Breiman, (2001)]
  - (intrinsic importance) ... by best possible (i.e., oracle) algorithm [e.g., van der Laan (2006)]
- Our work focuses on interpretable, model-agnostic intrinsic importance

Example uses of intrinsic variable importance:

- is it worth extracting text from notes in the EHR for the sake of predicting hospital readmission?

Variable importance: what and why

What is variable importance?

- Quantification of "contributions" of a variable (or a set of variables)

  Traditionally: contribution to predictions

- Useful to distinguish between contributions of predictions...
  - (extrinsic importance) ... by a given (possibly black-box) algorithm [e.g., Breiman, (2001)]
  - (intrinsic importance) ... by best possible (i.e., oracle) algorithm [e.g., van der Laan (2006)]
- Our work focuses on interpretable, model-agnostic intrinsic importance

Example uses of intrinsic variable importance:

- is it worth extracting text from notes in the EHR for the sake of predicting hospital readmission?
- does the importance of item 9 on the Patient Health Questionnaire in predicting risk of suicide attempt change over time?

Case study: ANOVA importance

Data unit $(X, Y) \sim P_0$ with:

- outcome Y
- covariate $X := (X_1, X_2, \ldots, X_p)$

Case study: ANOVA importance

Data unit $(X, Y) \sim P_0$ with:

- outcome Y
- covariate $X := (X_1, X_2, \ldots, X_p)$

Goals:

- estimate
- and do inference on

the importance of $(X_j : j \in s)$ in predicting Y

Case study: ANOVA importance

Data unit $(X, Y) \sim P_0$ with:

- outcome Y
- covariate $X := (X_1, X_2, \ldots, X_p)$

Goals:

- estimate
- and do inference on

the importance of $(X_j : j \in s)$ in predicting Y

How do we typically do this in **linear regression**?

Case study: ANOVA importance

How do we typically do this in linear regression?

- Fit a linear regression of Y on $X \rightarrow \mu_n(X)$

Case study: ANOVA importance

How do we typically do this in linear regression?

- Fit a linear regression of Y on $X \to \mu_n(X)$
- Fit a linear regression of Y on $X_{-s} \to \mu_{n,-s}(X)$

Case study: ANOVA importance

How do we typically do this in linear regression?

- Fit a linear regression of Y on $X \to \mu_n(X)$
- Fit a linear regression of Y on $X_{-s} \to \mu_{n,-s}(X)$
- Compare the fitted values $[\mu_n(X_i), \mu_{n,-s}(X_i)]$

Case study: ANOVA importance

How do we typically do this in linear regression?

- Fit a linear regression of Y on $X \to \mu_n(X)$
- Fit a linear regression of Y on $X_{-s} \to \mu_{n,-s}(X)$
- Compare the fitted values $[\mu_n(X_i), \mu_{n,-s}(X_i)]$

Many ways to compare fitted values, including:

- ANOVA decomposition
- Difference in R^2

Case study: ANOVA importance

Difference in R^2:

$$\left[1 - \frac{n^{-1}\sum_{i=1}^{n}\{Y_i - \mu_n(X_i)\}^2}{n^{-1}\sum_{i=1}^{n}\{Y_i - \overline{Y}_n\}^2}\right] - \left[1 - \frac{n^{-1}\sum_{i=1}^{n}\{Y_i - \mu_{n,-s}(X_i)\}^2}{n^{-1}\sum_{i=1}^{n}\{Y_i - \overline{Y}_n\}^2}\right]$$

Case study: ANOVA importance

Difference in R^2:

$$\left[1 - \frac{n^{-1}\sum_{i=1}^{n}\{Y_i - \mu_n(X_i)\}^2}{n^{-1}\sum_{i=1}^{n}\{Y_i - \overline{Y}_n\}^2}\right] - \left[1 - \frac{n^{-1}\sum_{i=1}^{n}\{Y_i - \mu_{n,-s}(X_i)\}^2}{n^{-1}\sum_{i=1}^{n}\{Y_i - \overline{Y}_n\}^2}\right]$$

Inference:

- Test difference
- Valid confidence interval

Case study: ANOVA importance

Consider the population parameter

$$\psi_{0,s} = \frac{E_0\{\mu_0(X) - \mu_{0,-s}(X)\}^2}{var_0(Y)}$$

- $\mu_0(x) := E_0(Y \mid X = x)$ (true conditional mean)
- $\mu_{0,-s}(x) := E_0(Y \mid X_{-s} = x_{-s})$

  [for a vector z, z_{-s} represents $(z_j : j \notin s)$]

Case study: ANOVA importance

Consider the population parameter

$$\psi_{0,s} = \frac{E_0\{\mu_0(X) - \mu_{0,-s}(X)\}^2}{var_0(Y)}$$

- $\mu_0(x) := E_0(Y \mid X = x)$ (true conditional mean)
- $\mu_{0,-s}(x) := E_0(Y \mid X_{-s} = x_{-s})$

  [for a vector z, z_{-s} represents $(z_j : j \notin s)$]
- nonparametric extension of linear regression-based ANOVA parameter

Case study: ANOVA importance

Consider the population parameter

$$\psi_{0,s} = \frac{E_0\{\mu_0(X) - \mu_{0,-s}(X)\}^2}{var_0(Y)}$$

- $\mu_0(x) := E_0(Y \mid X = x)$ (true conditional mean)
- $\mu_{0,-s}(x) := E_0(Y \mid X_{-s} = x_{-s})$

  [for a vector z, z_{-s} represents $(z_j : j \notin s)$]
- nonparametric extension of linear regression-based ANOVA parameter
- Can be expressed as a difference in population R^2 values, since

$$\psi_{0,s} = \left[1 - \frac{E_0\{Y - \mu_0(X)\}^2}{var_0(Y)}\right] - \left[1 - \frac{E_0\{Y - \mu_{0,-s}(X)\}^2}{var_0(Y)}\right]$$

Case study: ANOVA importance

How should we make inference on $\psi_{0,s}$?

Case study: ANOVA importance

How should we make inference on $\psi_{0,s}$?

1. construct estimators μ_n, $\mu_{n,-s}$ of μ_0 and $\mu_{0,-s}$ (e.g., with machine learning)

Case study: ANOVA importance

How should we make inference on $\psi_{0,s}$?

1. construct estimators μ_n, $\mu_{n,-s}$ of μ_0 and $\mu_{0,-s}$ (e.g., with machine learning)
2. plug in:

$$\psi_{n,s} := \frac{\frac{1}{n}\sum_{i=1}^{n}\{\mu_n(X_i) - \mu_{n,-s}(X_i)\}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}$$

Case study: ANOVA importance

How should we make inference on $\psi_{0,s}$?

1. construct estimators μ_n, $\mu_{n,-s}$ of μ_0 and $\mu_{0,-s}$ (e.g., with machine learning)
2. plug in:

$$\psi_{n,s} := \frac{\frac{1}{n}\sum_{i=1}^{n}\{\mu_n(X_i) - \mu_{n,-s}(X_i)\}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}$$

but this estimator has asymptotic bias

Case study: ANOVA importance

How should we make inference on $\psi_{0,s}$?

1. construct estimators μ_n, $\mu_{n,-s}$ of μ_0 and $\mu_{0,-s}$ (e.g., with machine learning)
2. plug in:

$$\psi_{n,s} := \frac{\frac{1}{n}\sum_{i=1}^{n}\{\mu_n(X_i) - \mu_{n,-s}(X_i)\}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}$$

   but this estimator has asymptotic bias
3. using influence function-based debiasing [e.g., Pfanzagl (1982)], we get estimator

$$\psi^*_{n,s} := \left[1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \mu_n(X_i)\}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}\right] - \left[1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \mu_{n,-s}(X_i)\}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}\right]$$

Case study: ANOVA importance

How should we make inference on $\psi_{0,s}$?

1. construct estimators μ_n, $\mu_{n,-s}$ of μ_0 and $\mu_{0,-s}$ (e.g., with machine learning)
2. plug in:

$$\psi_{n,s} := \frac{\frac{1}{n}\sum_{i=1}^{n}\{\mu_n(X_i) - \mu_{n,-s}(X_i)\}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}$$

   but this estimator has asymptotic bias

3. using influence function-based debiasing [e.g., Pfanzagl (1982)], we get estimator

$$\psi^*_{n,s} := \left[1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \mu_n(X_i)\}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}\right] - \left[1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \mu_{n,-s}(X_i)\}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}\right]$$

Under regularity conditions, $\psi^*_{n,s}$ is consistent and nonparametric efficient.

In particular, $\sqrt{n}(\psi^*_{n,s} - \psi_{0,s})$ has a mean-zero normal limit with estimable variance.

[Details in Williamson et al. (2020)]

Generalization to arbitrary measures

ANOVA example suggests a natural generalization:

Generalization to arbitrary measures

ANOVA example suggests a natural generalization:

- Choose a relevant measure of predictiveness for the task at hand

Generalization to arbitrary measures

ANOVA example suggests a natural generalization:

- Choose a relevant measure of predictiveness for the task at hand
  - $V(f, P)$ = predictiveness of function f under sampling from P
  - $\mathcal{F}$ = rich class of candidate prediction functions
  - $\mathcal{F}_{-s}$ = {all functions in $\mathcal{F}$ that ignore components with index in s} $\subset \mathcal{F}$

Generalization to arbitrary measures

ANOVA example suggests a natural generalization:

- Choose a relevant measure of predictiveness for the task at hand
  - $V(f, P) =$ predictiveness of function f under sampling from P
  - $\mathcal{F} =$ rich class of candidate prediction functions
  - $\mathcal{F}_{-s} = \{$all functions in $\mathcal{F}$ that ignore components with index in $s\} \subset \mathcal{F}$
- Define the oracle prediction functions

$f_0 :=$ maximizer of $V(f, P_0)$ over $\mathcal{F}$ & $f_{0,-s} :=$ maximizer of $V(f, P_0)$ over $\mathcal{F}_{-s}$

Generalization to arbitrary measures

ANOVA example suggests a natural generalization:

- Choose a relevant measure of predictiveness for the task at hand
  - $V(f, P) =$ predictiveness of function f under sampling from P
  - $\mathcal{F} =$ rich class of candidate prediction functions
  - $\mathcal{F}_{-s} = \{$all functions in $\mathcal{F}$ that ignore components with index in $s\} \subset \mathcal{F}$
- Define the oracle prediction functions

  $f_0 :=$ maximizer of $V(f, P_0)$ over $\mathcal{F}$ & $f_{0,-s} :=$ maximizer of $V(f, P_0)$ over $\mathcal{F}_{-s}$

Define the importance of $(X_j : j \in s)$ relative to X as

$$\psi_{0,s} := V(f_0, P_0) - V(f_{0,-s}, P_0) \geq 0$$

Generalization to arbitrary measures

Some examples of predictiveness measures:

(arbitrary outcomes)

R^2: $V(f, P) = 1 - E_P\{Y - f(X)\}^2 / var_P(Y)$

Generalization to arbitrary measures

Some examples of predictiveness measures:

(arbitrary outcomes)

R^2: $V(f,P) = 1 - E_P\{Y - f(X)\}^2 / var_P(Y)$

(binary outcomes)

Classification accuracy: $V(f,P) = P\{Y = f(X)\}$

AUC: $V(f,P) = P\{f(X_1) < f(X_2) \mid Y_1 = 0, Y_2 = 1\}$ for $(X_1, Y_1) \perp (X_2, Y_2)$

Pseudo- $R^2 : 1 - \frac{E_P[Y \log f(X) - (1-Y) \log\{1 - f(X)\}]}{P(Y=1) \log P(Y=1) + P(Y=0) \log P(Y=0)}$

Generalization to arbitrary measures

How should we make inference on $\psi_{0,s}$?

Generalization to arbitrary measures

How should we make inference on $\psi_{0,s}$?

1. construct estimators f_n, $f_{n,-s}$ of f_0 and $f_{0,-s}$ (e.g., with machine learning)

Generalization to arbitrary measures

How should we make inference on $\psi_{0,s}$?

1. construct estimators f_n, $f_{n,-s}$ of f_0 and $f_{0,-s}$ (e.g., with machine learning)
2. plug in:

$$\psi^*_{n,s} := V(f_n, P_n) - V(f_{n,-s}, P_n)$$

   where P_n is the empirical distribution based on the available data

Generalization to arbitrary measures

How should we make inference on $\psi_{0,s}$?

1. construct estimators f_n, $f_{n,-s}$ of f_0 and $f_{0,-s}$ (e.g., with machine learning)
2. plug in:

$$\psi^*_{n,s} := V(f_n, P_n) - V(f_{n,-s}, P_n)$$

   where P_n is the empirical distribution based on the available data
3. Inference can be carried out using influence functions.

Generalization to arbitrary measures

How should we make inference on $\psi_{0,s}$?

1. construct estimators f_n, $f_{n,-s}$ of f_0 and $f_{0,-s}$ (e.g., with machine learning)
2. plug in:

$$\psi^*_{n,s} := V(f_n, P_n) - V(f_{n,-s}, P_n)$$

   where P_n is the empirical distribution based on the available data

3. Inference can be carried out using influence functions. Why?

We can write
$V(f_n, P_n) - V(f_0, P_0) \approx V(f_0, P_n) - V(f_0, P_0) + V(f_n, P_0) - V(f_0, P_0)$

Generalization to arbitrary measures

How should we make inference on $\psi_{0,s}$?

1. construct estimators f_n, $f_{n,-s}$ of f_0 and $f_{0,-s}$ (e.g., with machine learning)

2. plug in:

$$\psi^*_{n,s} := V(f_n, P_n) - V(f_{n,-s}, P_n)$$

   where P_n is the empirical distribution based on the available data

3. Inference can be carried out using influence functions. Why?

We can write
$V(f_n, P_n) - V(f_0, P_0) \approx \color{green}{V(f_0, P_n) - V(f_0, P_0)} + \color{blue}{V(f_n, P_0) - V(f_0, P_0)}$

- the $\color{green}{\text{green term}}$ can be studied using the functional delta method
- the $\color{blue}{\text{blue term}}$ is second-order because f_0 maximizes V over $\mathcal{F}$

Generalization to arbitrary measures

How should we make inference on $\psi_{0,s}$?

1. construct estimators f_n, $f_{n,-s}$ of f_0 and $f_{0,-s}$ (e.g., with machine learning)
2. plug in:

$$\psi^*_{n,s} := V(f_n, P_n) - V(f_{n,-s}, P_n)$$

   where P_n is the empirical distribution based on the available data

3. Inference can be carried out using influence functions. Why?

We can write
$V(f_n, P_n) - V(f_0, P_0) \approx V(f_0, P_n) - V(f_0, P_0) + V(f_n, P_0) - V(f_0, P_0)$

- the green term can be studied using the functional delta method
- the blue term is second-order because f_0 maximizes V over $\mathcal{F}$

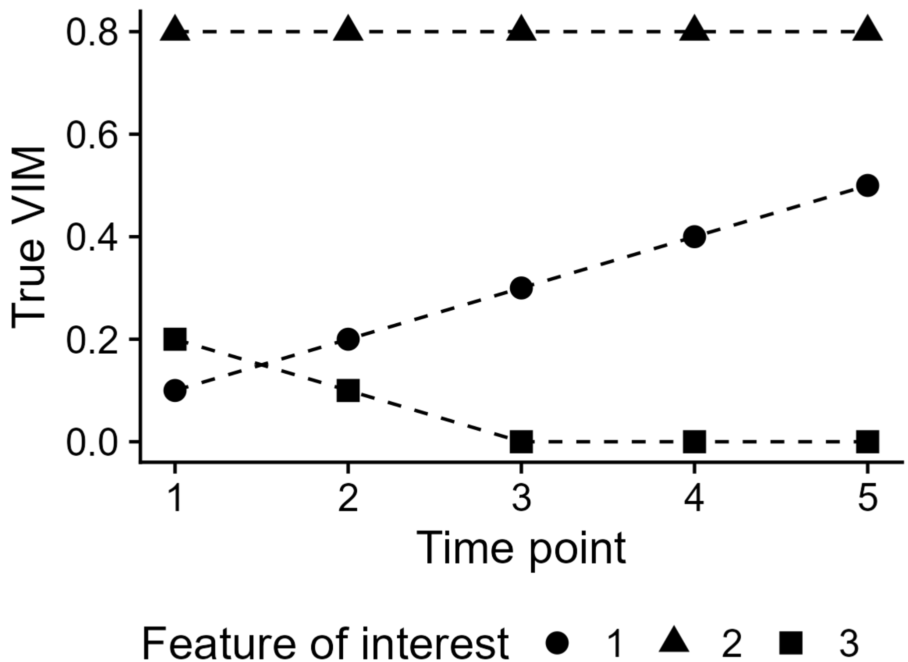
In other words: f_0 and $f_{0,-s}$ can be treated as known in studying behavior of $\psi^*_{n,s}$!

[Details in Williamson et al. (2022)]

Longitudinal VIMs

So far: cross-sectional variable importance

Can we do inference on variable importance longitudinally?



Summarizing a VIM trajectory

Define:

- contiguous set of timepoints $\tau := [t_0, t_1]$
- variable importance at each time point $\psi_{0,s,t}$, $t \in \tau$

Summarizing a VIM trajectory

Define:

- contiguous set of timepoints $\tau := [t_0, t_1]$
- variable importance at each time point $\psi_{0,s,t}$, $t \in \tau$

Examples of summary measures over τ, $m(\psi_{0,s,\tau})$:

Mean: $\|\tau\|^{-1} \sum_{t \in \tau} \psi_{0,s,t}$

Summarizing a VIM trajectory

Define:

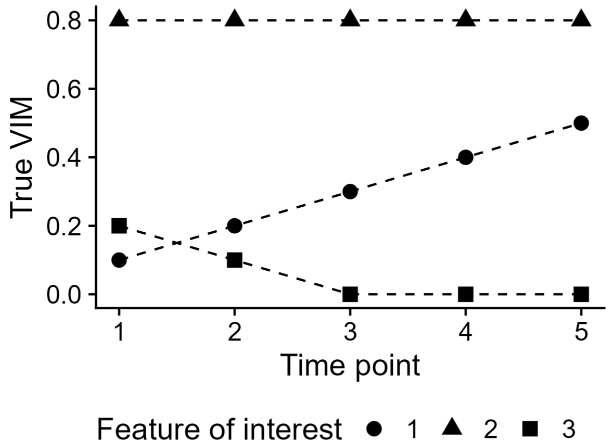
- contiguous set of timepoints $\tau := [t_0, t_1]$
- variable importance at each time point $\psi_{0,s,t}$, $t \in \tau$

Examples of summary measures over τ, $m(\psi_{0,s,\tau})$:

Mean: $\|\tau\|^{-1} \sum_{t \in \tau} \psi_{0,s,t}$

Linear trend: $(\beta_0, \beta_1) = \mathbf{arg\ min}_{(\alpha_1, \alpha_2) \in \mathrm{R}^2} \|\psi_{0,s,t} - \alpha_1 - t\alpha_2\|_2^2$

Summarizing a VIM trajectory



Summary	VIM 1	VIM 2	VIM 3
Mean	0.3	0.8	0.06
Slope	0.1	0	-0.05

Summarizing a VIM trajectory

How should we make inference on $m(\psi_{0,s,\tau})$?

Summarizing a VIM trajectory

How should we make inference on $m(\psi_{0,s,\tau})$?

1. construct estimators $f_{n,t}$, $f_{n,-s,t}$ of $f_{0,t}$, $f_{0,-s,t}$ (e.g., with machine learning)

Summarizing a VIM trajectory

How should we make inference on $m(\psi_{0,s,\tau})$?

1. construct estimators $f_{n,t}$, $f_{n,-s,t}$ of $f_{0,t}$, $f_{0,-s,t}$ (e.g., with machine learning)
2. plug in: $\psi^*_{n,s,t} = V(f_{n,s,t}, P_{n,s,t}) - V(f_{n,-s,t}, P_{n,s,t})$

Summarizing a VIM trajectory

How should we make inference on $m(\psi_{0,s,\tau})$?

1. construct estimators $f_{n,t}$, $f_{n,-s,t}$ of $f_{0,t}$, $f_{0,-s,t}$ (e.g., with machine learning)
2. plug in: $\psi^*_{n,s,t} = V(f_{n,s,t}, P_{n,s,t}) - V(f_{n,-s,t}, P_{n,s,t})$
3. plug in:

$$m^*_{n,s} := m(\psi^*_{n,s,\tau})$$

Summarizing a VIM trajectory

How should we make inference on $m(\psi_{0,s,\tau})$?

1. construct estimators $f_{n,t}$, $f_{n,-s,t}$ of $f_{0,t}$, $f_{0,-s,t}$ (e.g., with machine learning)
2. plug in: $\psi^*_{n,s,t} = V(f_{n,s,t}, P_{n,s,t}) - V(f_{n,-s,t}, P_{n,s,t})$
3. plug in:

$$m^*_{n,s} := m(\psi^*_{n,s,\tau})$$

4. Inference can be carried out using influence functions.

[Details in Williamson et al. (2023)]

VIMs for predictors of suicide risk

Data gathered from electronic health record on sample of 343,950 visits made by 184,782 people

Key variables: Jacobs et al. (2010)

- Patient Health Questionnaire (PHQ)
    - PHQ-8 total score (depressive symptoms)
    - PHQi9 (suicidal ideation)
- Prior recorded self-harm
- Age
- Sex (sex assigned at birth)

Outcome: suicide attempt in 90 days following mental health visit

Sampled one visit per person at six possible measurement times over 18 months:

- 99,991 people had only one visit
- 4,093 people had six visits
- Rate of suicide attempt approximately 0.5% at all time points

VIMs for predictors of suicide risk

Goal: estimate VIMs for PHQi9 and prior recorded self-harm

Variable sets considered:

1. no variables
2. PHQi9 alone
3. age and sex (base set)
4. age, sex, and PHQi9
5. age, sex, and prior self-harm
6. age, sex, prior self-harm, and PHQi9

VIMs for predictors of suicide risk

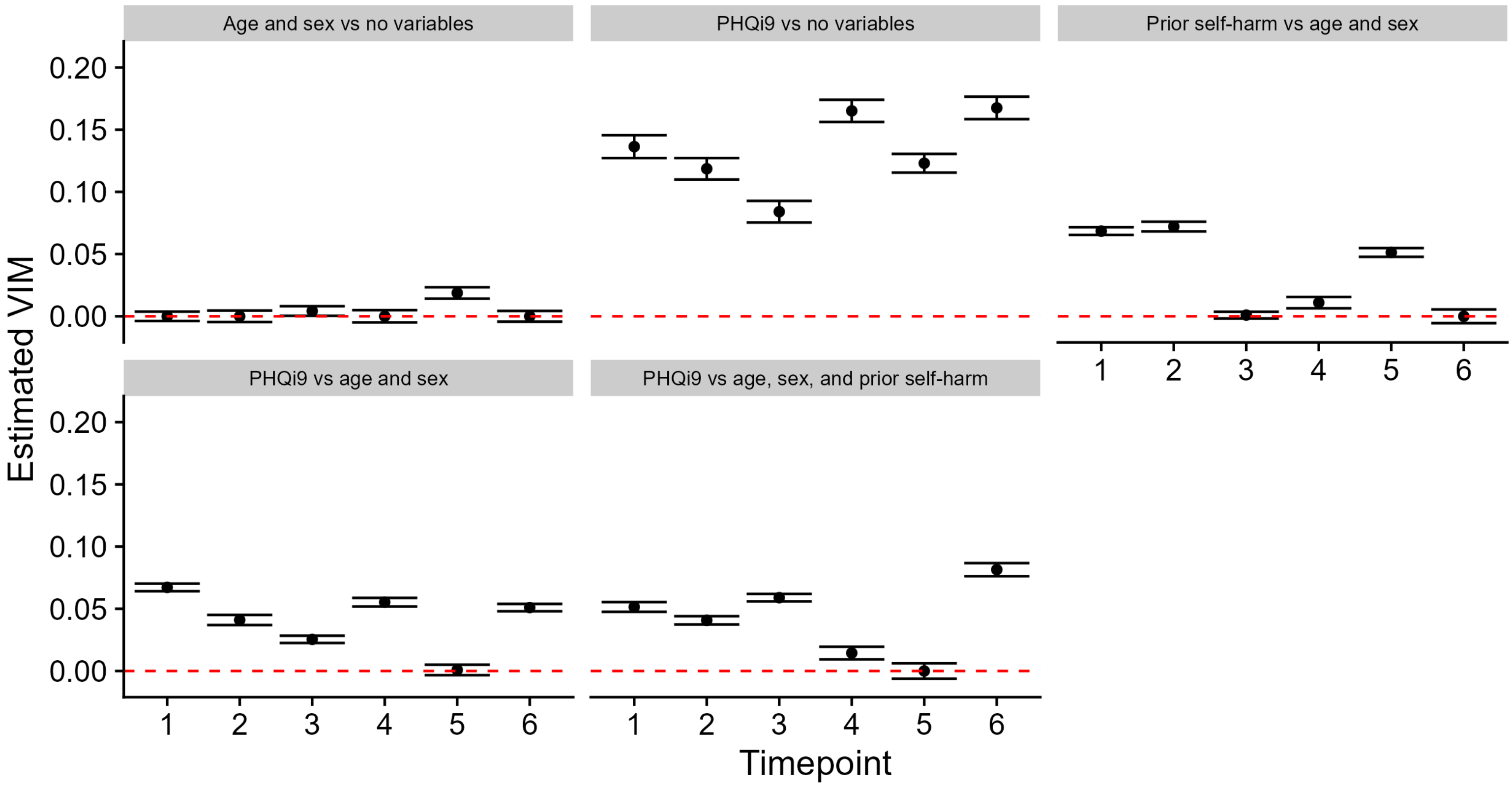
Goal: estimate VIMs for PHQi9 and prior recorded self-harm

Variable sets considered:

1. no variables
2. PHQi9 alone
3. age and sex (base set)
4. age, sex, and PHQi9
5. age, sex, and prior self-harm
6. age, sex, prior self-harm, and PHQi9

Estimate prediction functions at each time point using the Super Learner [van der Laan et al. (2007)]

VIMs for predictors of suicide risk



VIMs for predictors of suicide risk

Summary	Comparison	Estimate	SE	95% CI	p-value
Mean	PHQi9 versus no variables	0.132	0.019	[0.096, 0.169]	< 0.001
Mean	PHQi9 versus age and sex	0.040	0.012	[0.016, 0.064]	< 0.001
Mean	PHQi9 versus age, sex, and prior self-harm variables	0.033	0.012	[0.010, 0.056]	0.002
Trend: slope	PHQi9 versus no variables	0.007	0.011	[-0.014, 0.028]	0.507
Trend: slope	PHQi9 versus age and sex	-0.005	0.007	[-0.019, 0.009]	0.486
Trend: slope	PHQi9 versus age, sex, and prior self-harm variables	-0.005	0.007	[-0.018, 0.008]	0.468

Closing thoughts

Population-based variable importance:

- wide variety of meaningful measures
- simple estimators
- machine learning okay
- valid inference, testing
- extension to longitudinal VIMs
- extension to correlated features (Williamson and Feng, 2020)

Check out the software:

- R packages `vimp`, `lvimp`
- Python package `vimpy`

https://github.com/bdwilliamson | https://bdwilliamson.github.io

References

- Breiman L. 2001. Random forests. *Machine Learning*.
- van der Laan MJ. 2006. Statistical inference for variable importance. *The International Journal of Biostatistics*.
- van der Laan MJ, Polley EC, and Hubbard AE. 2007. Super Learner. *Statistical Applications in Genetics and Molecular Biology*.
- Williamson BD, Gilbert P, Carone M, and Simon N. 2020. Nonparametric variable importance assessment using machine learning techniques (+ rejoinder to discussion). *Biometrics*.
- Williamson BD, Gilbert P, Simon N, and Carone M. 2022. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*.
- Williamson BD, Moodie EEM, and Shortreed SM. 2023. Inference on summaries of a model-agnostic longitudinal variable importance trajectory. *arXiv* https://arxiv.org/pdf/2311.01638.pdf.
- Williamson BD and Feng J. 2020. Efficient nonparametric statistical inference on population feature importance using Shapley values. *ICML*.