LOCO: The Good, the Bad, and the Ugly

(or: How I Learned to Stop Worrying and Love Prediction)

Ryan Tibshirani
Depts. of Statistics & Machine Learning
Carnegie Mellon University

Thanks to: Jing Lei, Max G'Sell, Alessandro Rinaldo, Larry Wasserman, Jon Taylor, Rob Tibshirani

`http://www.stat.cmu.edu/~ryantibs/talks/loco-2018.pdf`

What can we do without a model?

Given i.i.d. (X^i, Y^i), $i = 1, \ldots, n$, from a distribution P on $\mathbb{R}^d \times \mathbb{R}$. Denote $f(x) = \mathbb{E}(Y|X = x)$

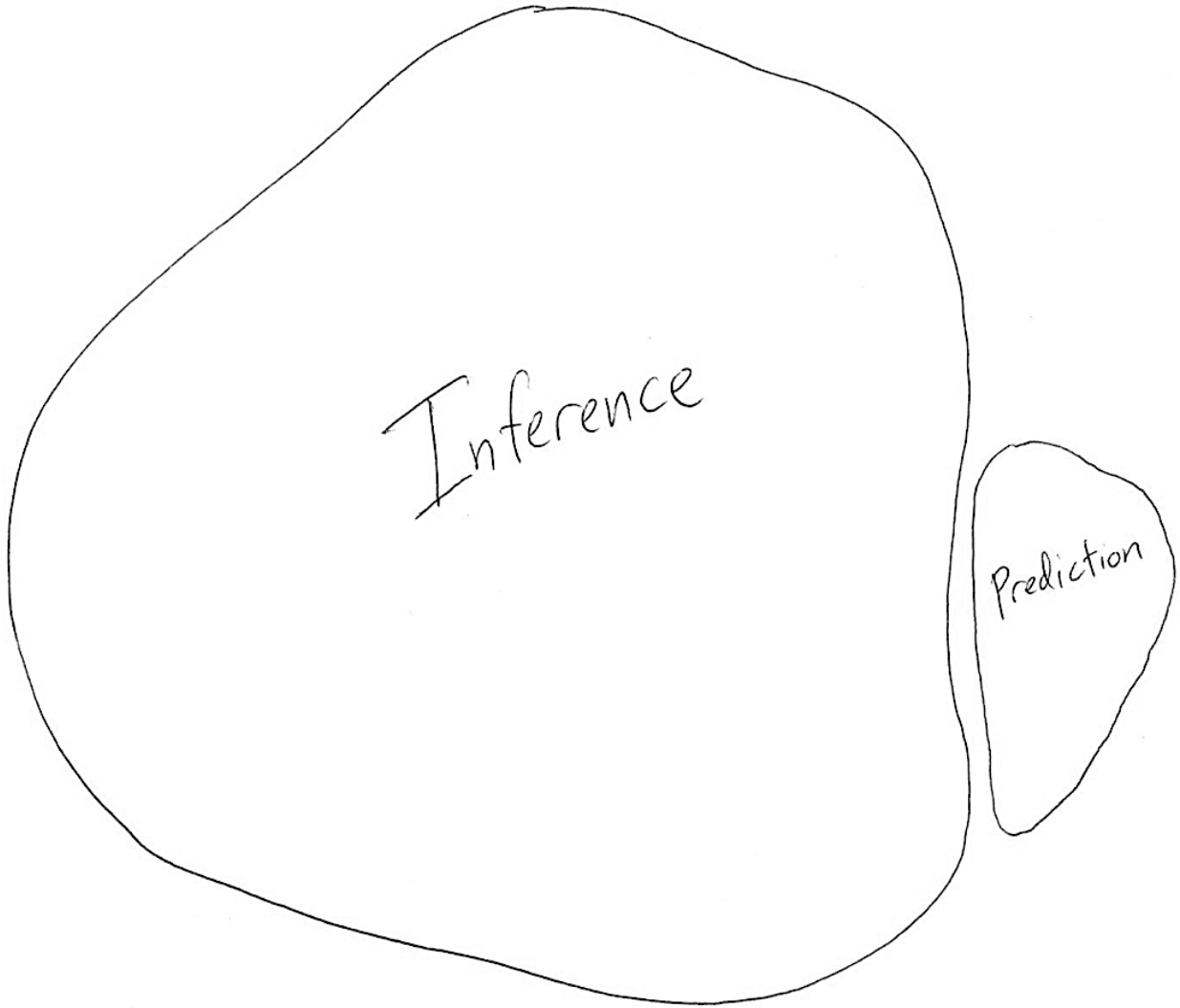
Main question: without assuming model for P (or for $P_{Y|X}$), what can we say in terms of variable importance? For selected variables?

Depends if we care about testing hypotheses or covering parameters (p-values versus confidence intervals)

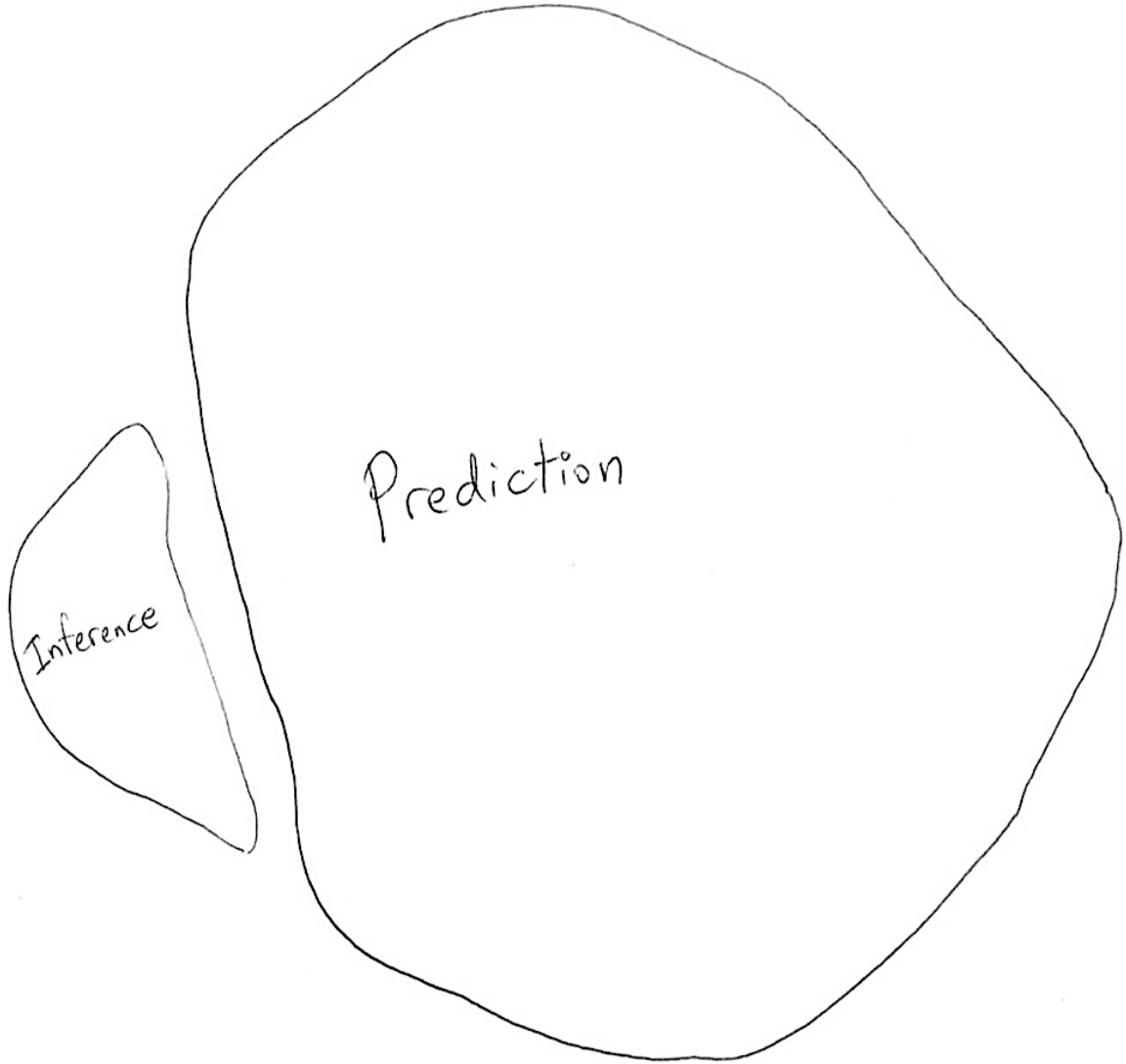
- Many interesting model-free hypothesis tests ...
- But model-free parameters? (I.e., functionals?)

Oxymoron aside, this is an important practical question ... often we are interested in effect sizes. Main question, rephrased: are there interesting functionals that measure effect sizes in model-free way?

Stats view of the world



ML view of the world



Other approaches?

- Projection parameter: for some class of functions f_θ, "working model" (typically parametric, e.g., $f_\theta(x) = \theta^T x$), consider

$$\theta^* = \arg\min_\theta \ \mathbb{E}\big[Y - f_\theta(X)\big]^2$$

  Studied by Larry's group, Jon's group, etc. Do inferences on f_{θ^*}. Somewhat limiting?

- Distance-to-independence: for some choice of distance d (e.g., KL, TV, Wasserstein), consider

$$d\Big(P(X_j, Y|X_{-j}),\ P(X_j|X_{-j}) \times P(Y|X_{-j})\Big)$$

  How to cover this? Lacks interpretability?

Outline

- The LOCO approach
- The good/bad/ugly
- An algorithm-free target
- Purely predictive variant
- Conclusions

The LOCO approach

LOCO inference

In Lei et al. (2016), we proposed a simple idea for measuring variable importance, called leave-one-covariate-out (LOCO) inference:

- Split samples into two parts, $D_1 \cup D_2 = \{1, \ldots n\}$
- Run algorithm to compute estimate $\hat{f}_{n_1}$ on first part D_1
- Select some interesting variable j, recompute $\hat{f}_{n_1}^{-j}$ on first part (rerun algorithm without access to variable j)
- Use second part D_2 to construct finite-sample, distribution-free confidence interval (e.g., use sign test or Wilcoxon test) for

$$\theta_j(D_1) = \mathrm{med}\Big(|Y - \hat{f}_{n_1}^{-j}(X)| - |Y - \hat{f}_{n_1}(X)| \,\Big|\, D_1\Big)$$

Note: algorithm for estimating $f(x) = \mathbb{E}(Y|X = x)$ can be arbitrary (lasso, lasso + CV, random forest, gradient boosting, neural net ...)

Examples of LOCO

Example: $n = 200$, $d = 500$, with data model $Y = \beta^T X + \epsilon$, such that $X \sim N(0, I_d)$, $\epsilon \sim N(0, 1)$, and

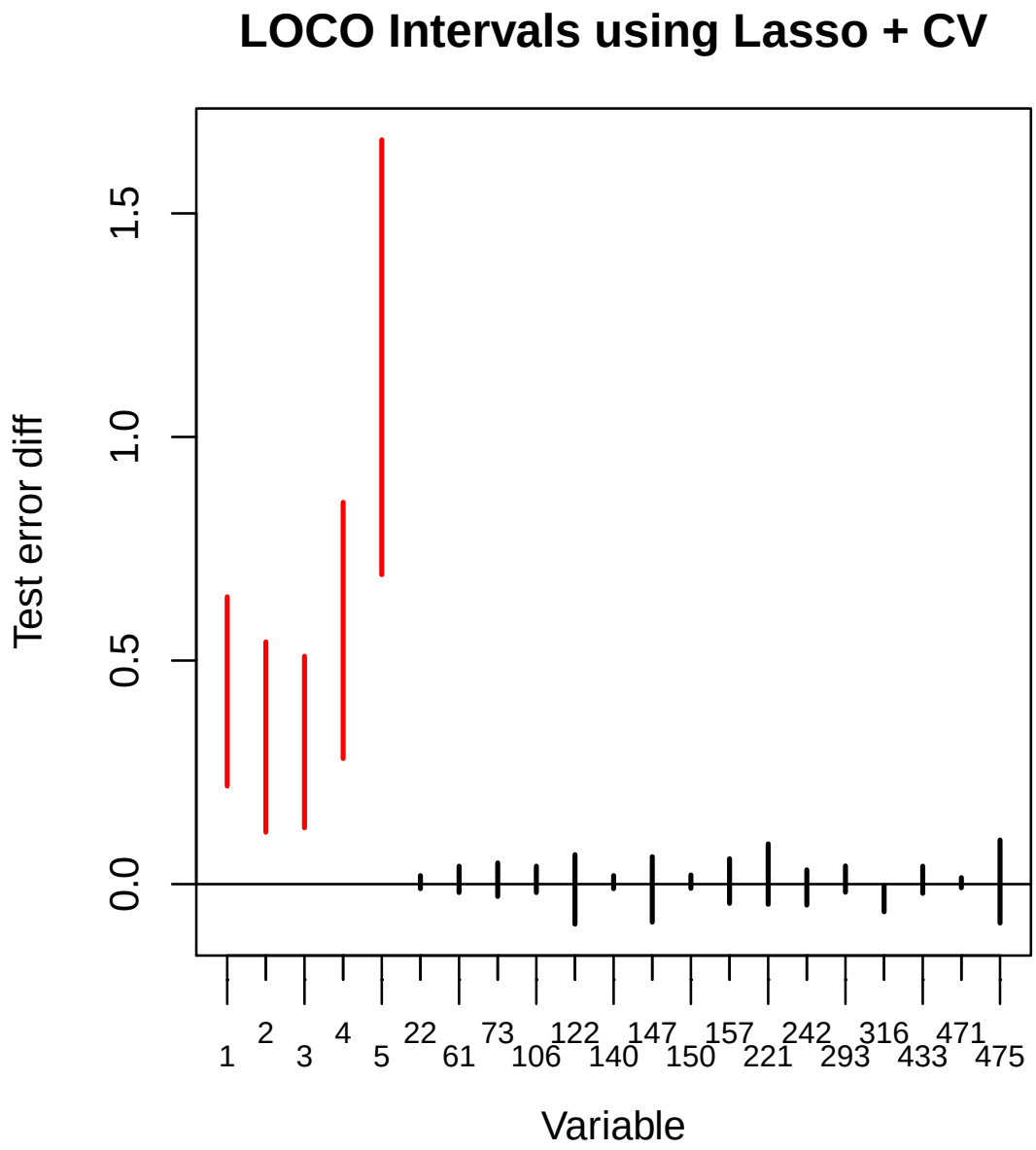
$$\beta_j \begin{cases} \sim N(0, 2) & j = 1, \ldots, 5 \\ = 0 & \text{otherwise} \end{cases}$$

- Algorithm is the lasso, with 5-fold CV and 1se rule to select λ
- Compute an interval for

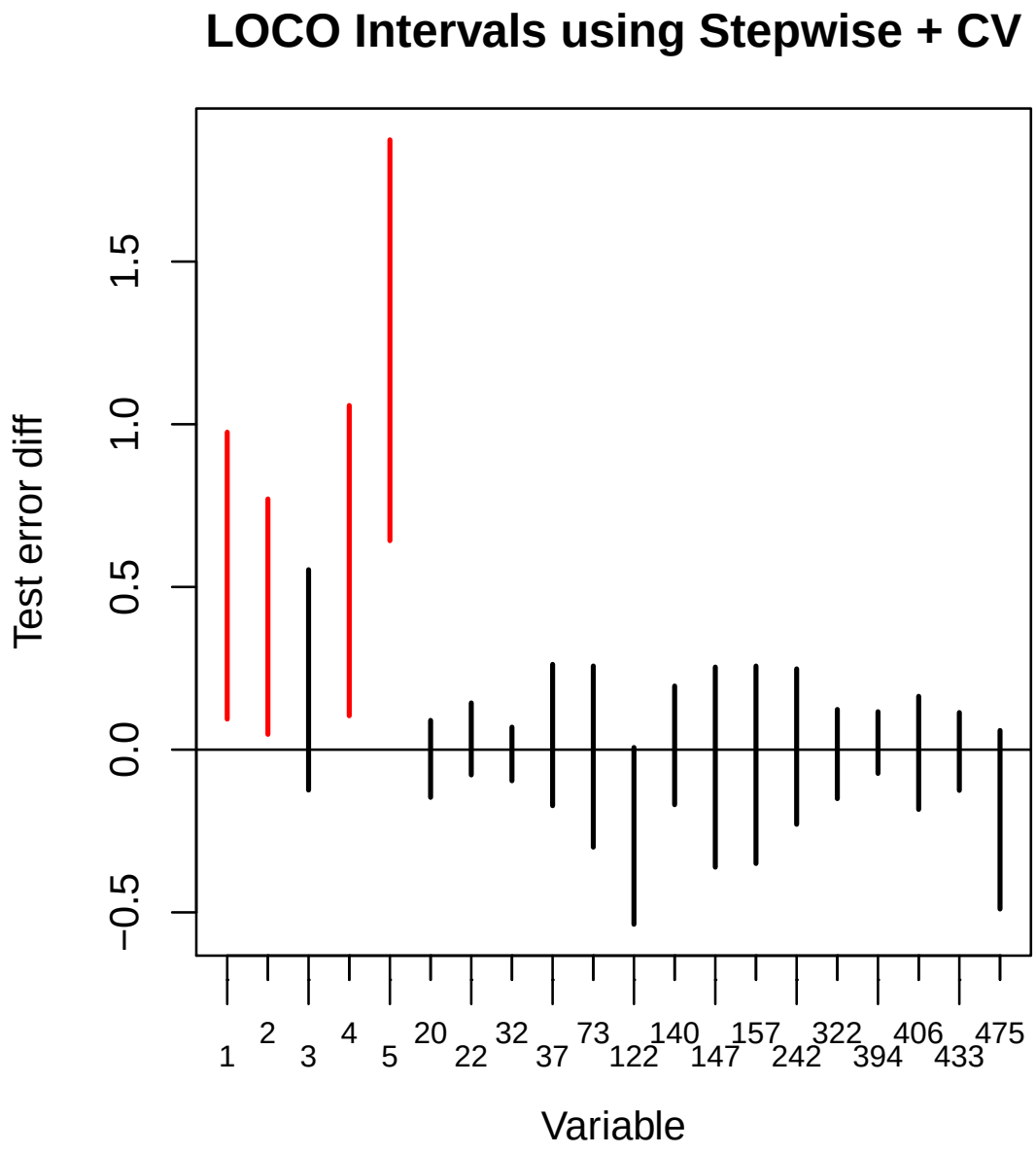
$$\theta_j(D_1) = \text{med}\Big(|Y - \hat{f}_{n_1}^{-j}(X)| - |Y - \hat{f}_{n_1}(X)| \,\Big|\, D_1 \Big)$$

  for each j in lasso active set
- Use Bonferroni correction: if s variables are selected, then we compute each LOCO interval at level $1 - \alpha/s$

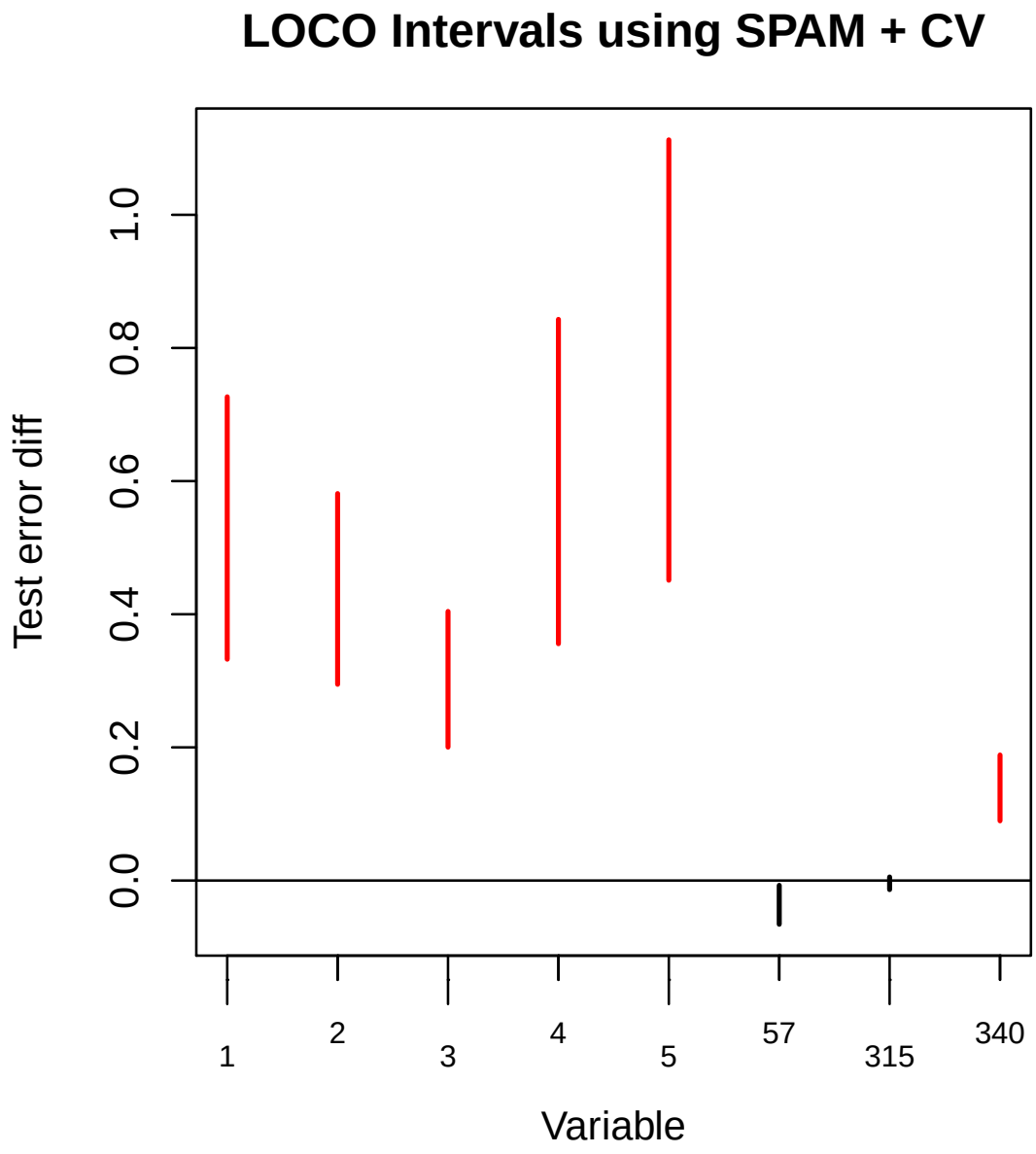
Examples of LOCO



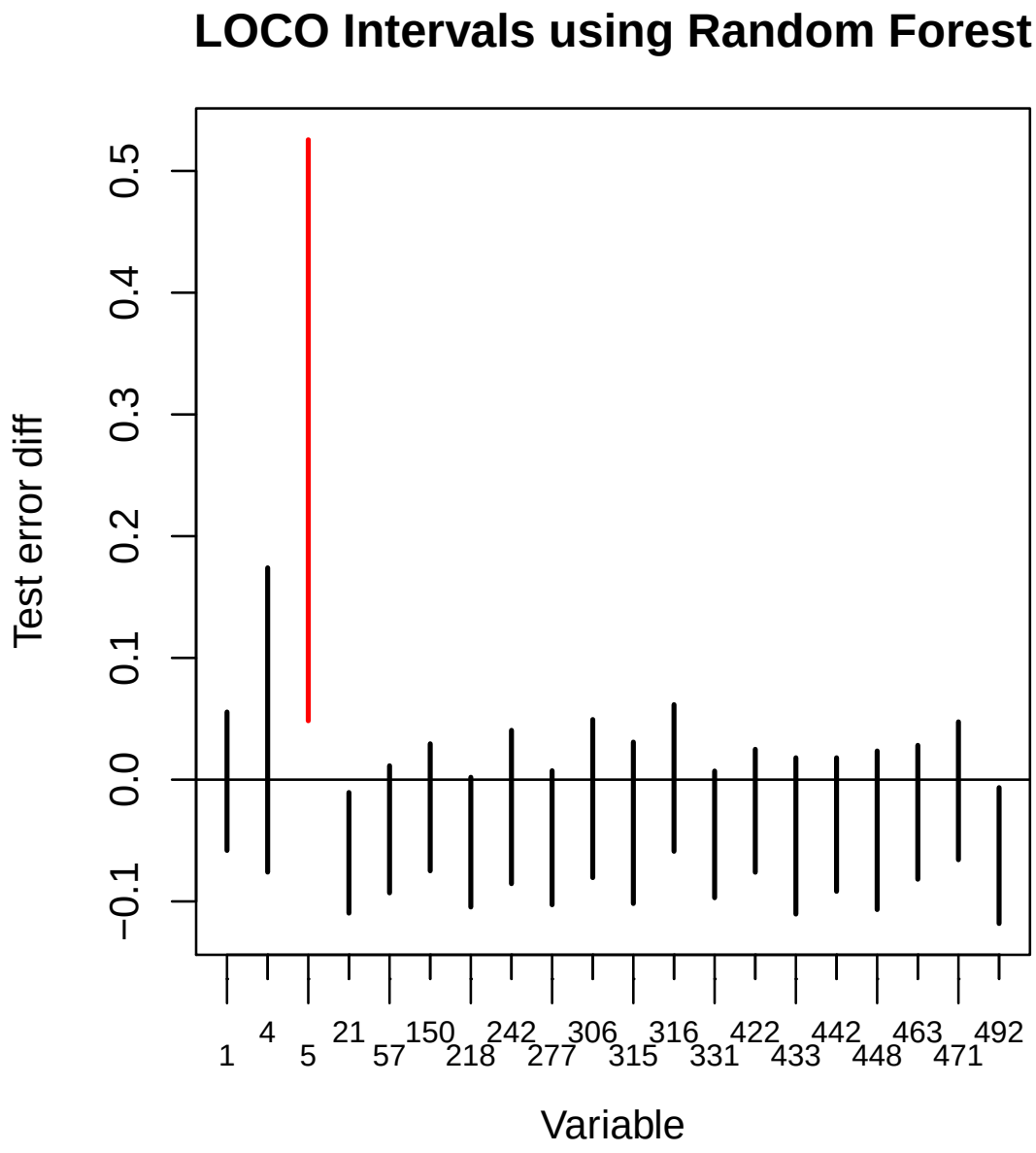
Examples of LOCO



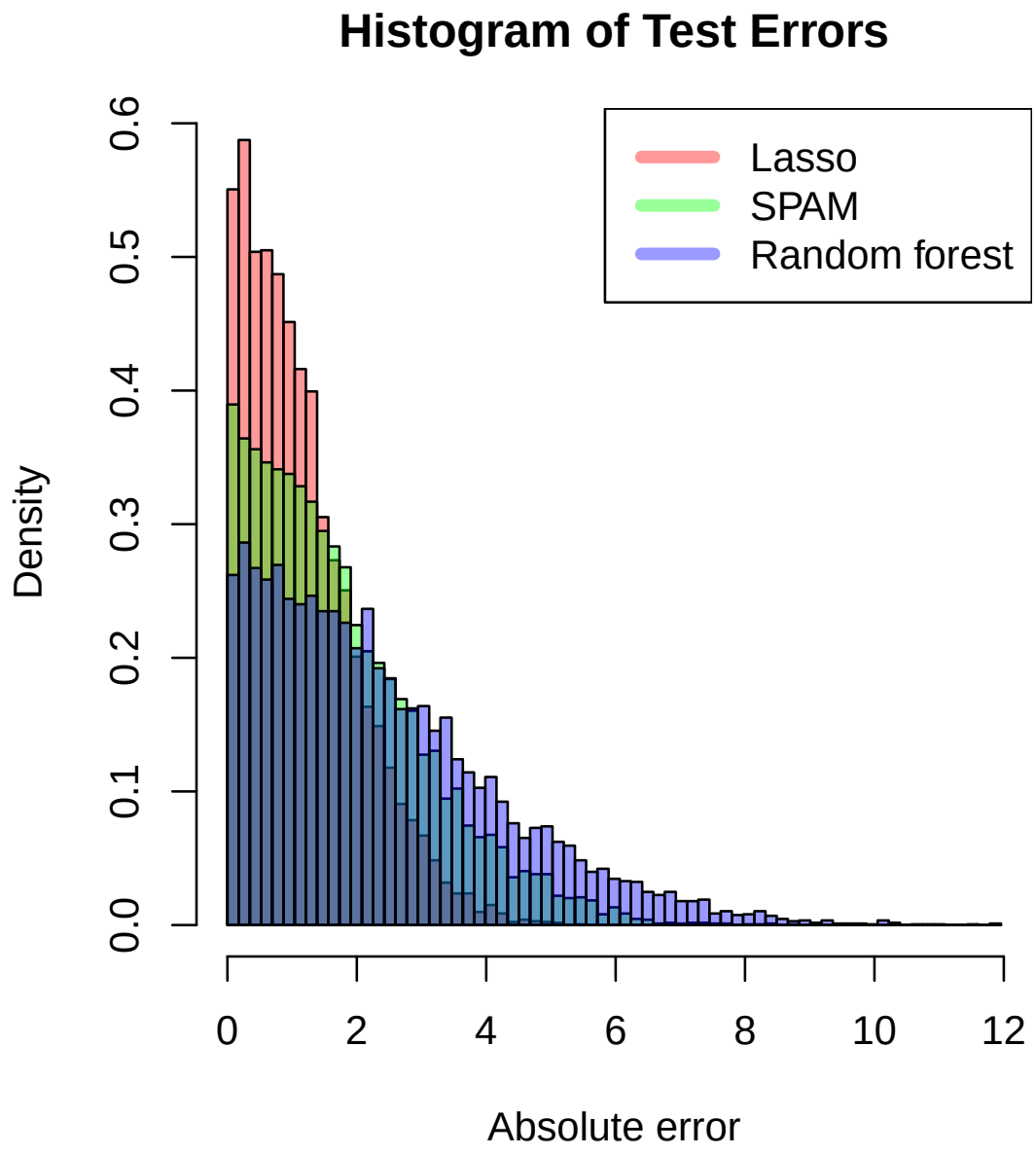
Examples of LOCO



Examples of LOCO



Examples of LOCO



The good/bad/ugly

The good

- Algorithmically flexible: any algorithm can be used to measure variable importance
- Computationally cheap(-ish): one refitting of the algorithm at hand per variable considered
- No distributional assumptions: intervals for $\theta_j(D_1)$ have exact coverage in finite-sample, for any distribution P of (X, Y)
- Selective validity: intervals cover the selected variables
- Accuracy: Rinaldo et al. (2016) show intervals (with Bonferroni correction, for s variables) have length $O(\sqrt{\log(sn)/n})$
- Simplicity: very simple/portable! Easy implementation

The bad

- The LOCO parameter is not on an intuitive scale
- Results we declare in practice are tied to choice of algorithm
- Results are also sensitive to ratio of training to test set sizes

Fixes.

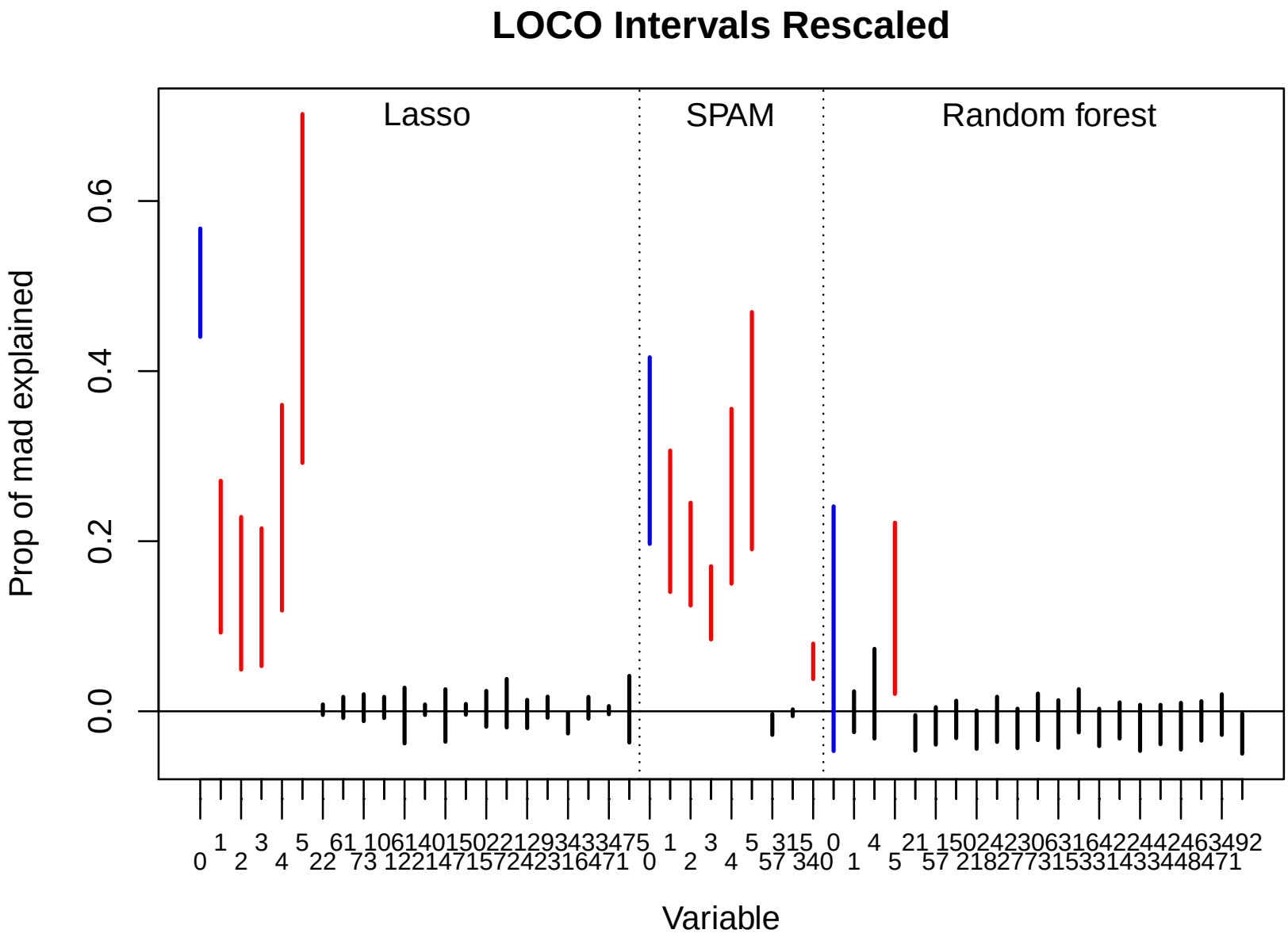
  - Rescale LOCO parameter

$$\theta_j(D_1) = \frac{\text{med}\left(|Y - \hat{f}_{n_1}^{-j}(X)| - |Y - \hat{f}_{n_1}(X)| \,\middle|\, D_1\right)}{\text{mad}(Y)}$$

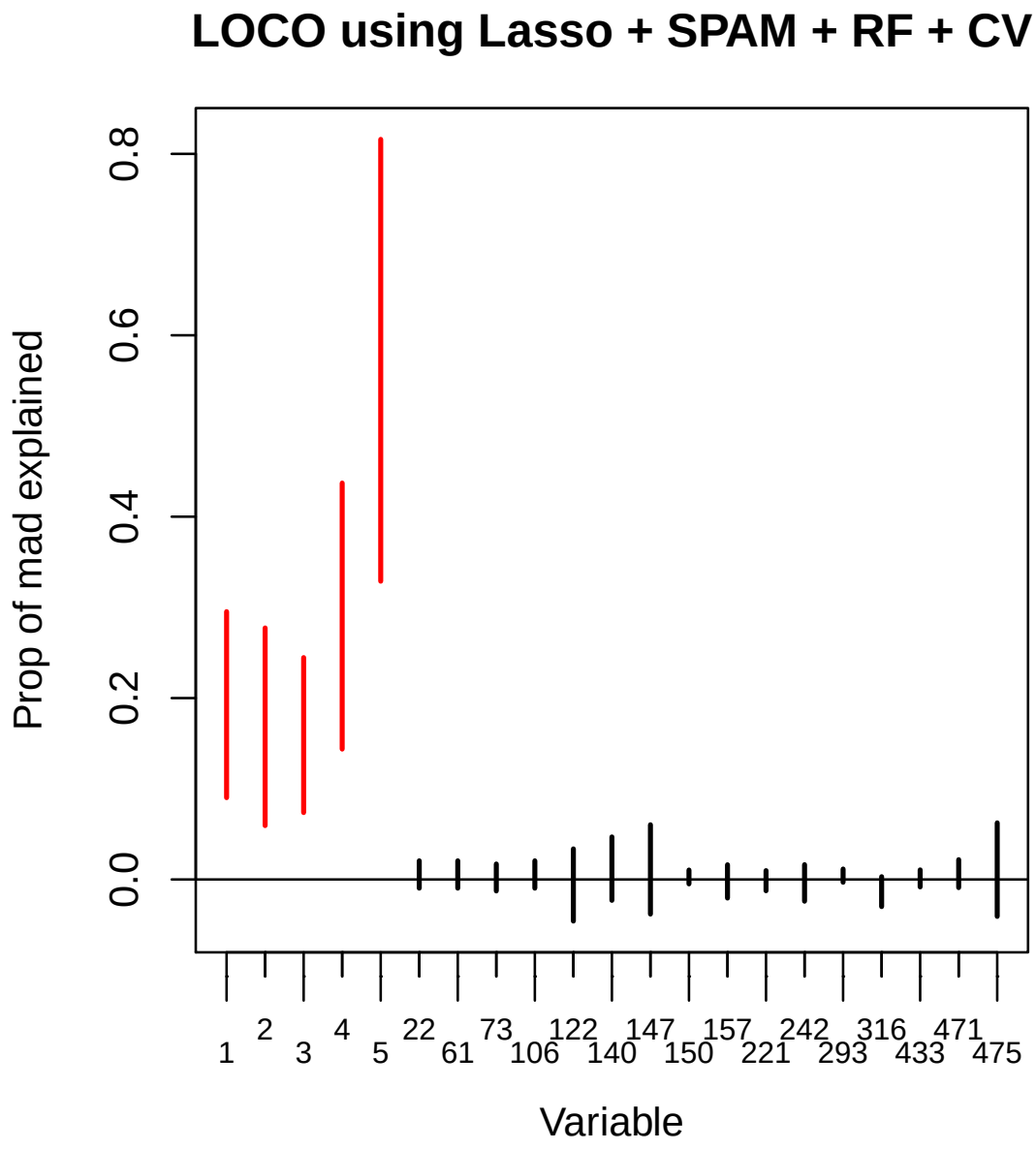
  - In defining algorithm, use something like CV (or meta-CV)
  - Cover both LOCO parameter and

$$\theta_0(D_1) = \frac{\text{med}\left(|Y - \hat{f}_{n_1}(X)| - \text{mad}(Y) \,\middle|\, D_1\right)}{\text{mad}(Y)}$$

Examples of LOCO



Examples of LOCO



The ugly

(aside from the name)

The parameter

$$\theta_j(D_1) = \text{med}\Big(|Y - \hat{f}_{n_1}^{-j}(X)| - |Y - \hat{f}_{n_1}(X)| \,\Big|\, D_1\Big)$$

is conditional on D_1. It measures "how important is variable j, to our *algorithm's estimates on* D_1?"

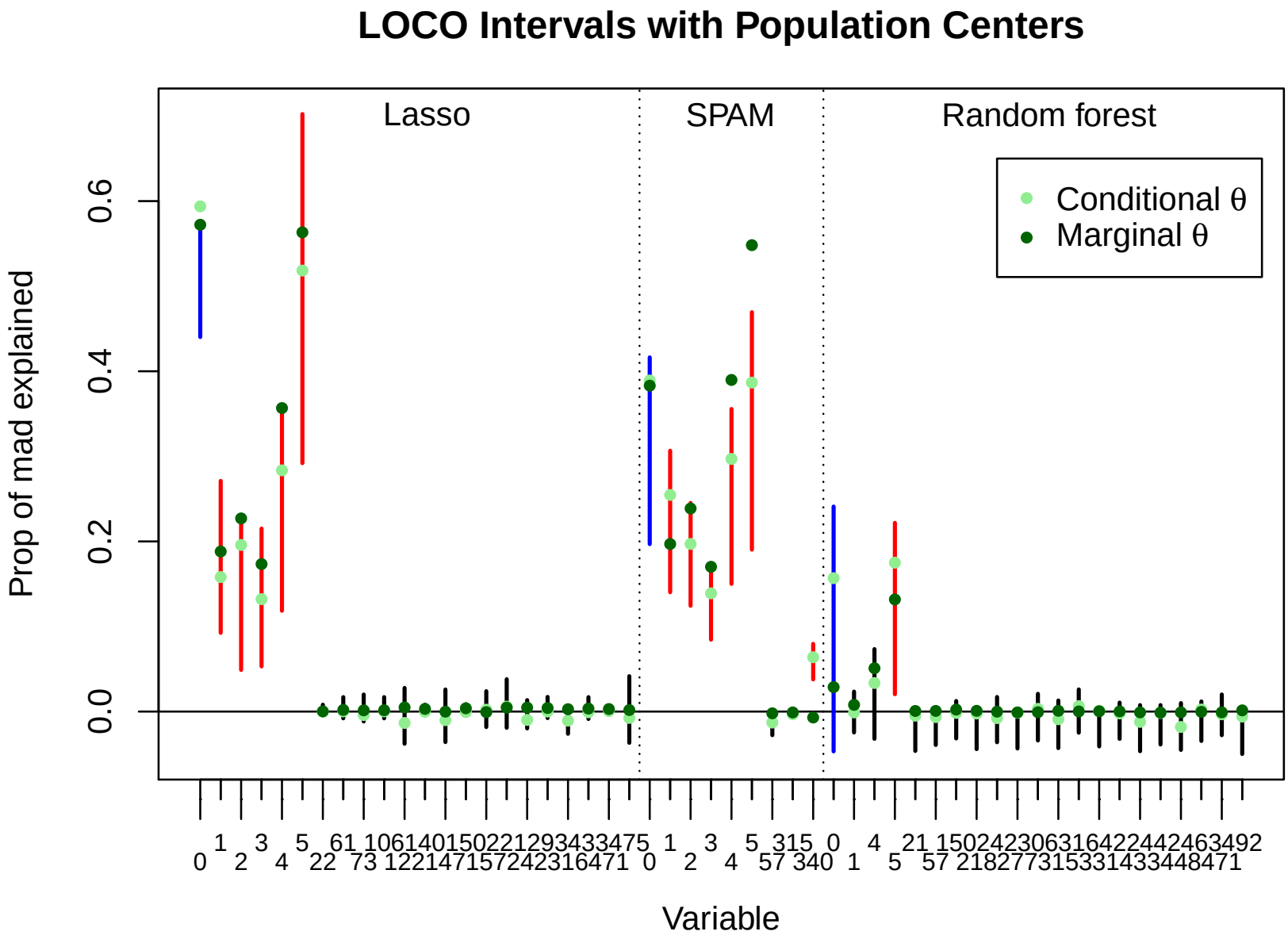
Compare this to

$$\theta_j = \text{med}\Big(|Y - \hat{f}_{n_1}^{-j}(X)| - |Y - \hat{f}_{n_1}(X)|\Big)$$

which measures "how important is variable j, to our *algorithm run on* n_1 *samples*?"

These are not the same!

Examples of LOCO



Marginal or conditional?

- Parameter θ_j itself is (arguably) more natural/interesting
- Asymptotically, interval for $\theta_j(D_1)$ need not be centered around θ_j, even in simple settings (Taylor, personal communication)
- Generic multi-splitting won't work either—each time we cover a different parameter, and not clear how to combine inferences
- Multi-splitting *can* work if we take small training sets, adopt a U-statistic view, but now parameter has very different meaning
- Markovic, Xia, Taylor (2017) cover θ_j using randomization + normal approximation ... requires assumptions
- Distribution-free inference for θ_j is an open problem

An algorithm-free target

Sans algorithm

Absent of any algorithm, we can still consider

$$\phi_j = \frac{\|Y - \mathbb{E}(Y|X_{-j})\|^2 - \|Y - \mathbb{E}(Y|X)\|^2}{\mathrm{Var}(Y)}$$

(where $\|Z\|^2 = \mathbb{E}(Z^2)$). This is like a nonparametric proportion of variance explained by X_j.

Equivalent form:

$$\phi_j = \frac{\|\mathbb{E}(Y|X_{-j}) - \mathbb{E}(Y|X)\|^2}{\mathrm{Var}(Y)}$$

We could also study more robust version (expectations $\to$ medians)

Study ϕ_j or θ_j?

LOCO was initially motivated by (more robust version of):

$$\theta_j = \frac{\|Y - \hat{\mathbb{E}}(Y|X_{-j})\|^2 - \|Y - \hat{\mathbb{E}}(Y|X)\|^2}{\mathrm{Var}(Y)}$$

for particular plug-in estimators of the conditional expectations

This was done to get distribution-free, finite-sample results. But we could also motivate our study by ϕ_j, which is algorithm-free

- Problem is we must ask for consistency, resort to asymptotics
- Williamson et al. (2017) show how to do inference for ϕ_j using semiparametric theory ... requires assumptions
- Again, assumption-lean inference for ϕ_j is an open problem

Purely predictive variant

Predictive perspective

Consider the random variable:

$$\Delta_j(X,Y) = |Y - \hat{f}_n^{(-j)}(X)| - |Y - \hat{f}_n(X)|$$

where $\hat{f}_n$, $\hat{f}_n^{(-j)}$ are fit on full data set, and (X,Y) is a new pair

Somewhat remarkably, using theory of conformal inference, we can get a distribution-free, finite-sample prediction band for $\Delta_j(X,Y)$:

$$\mathbb{P}\Big(\Delta_j(X,Y) \in C_j(X),\ j = 1,\ldots,d\Big) \geq 1 - \alpha$$

Note the simultaneity over all variables j! Two important notes:

- Coverage is marginal over $\hat{f}_n$, $\hat{f}_n^{(-j)}$
- Coverage is also marginal over X

(Could this give an avenue for assumption-lean inference on θ_j?)

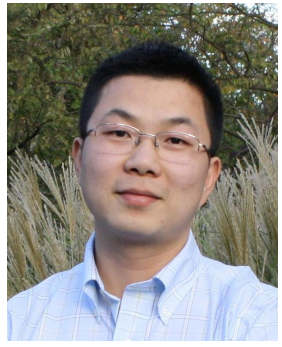
Conclusions

Summary

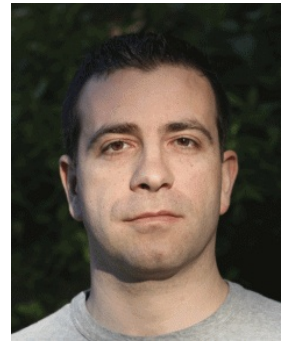
- Inference on effect sizes for variable importance in a model-free way is an important problem
- LOCO provides a fast, simple, distribution-free coverage of the median excess test error after omitting any given variable
- We can put target on a natural scale: prop of mad explained, weaken practical dependence on algorithm, and interpret in an algorithm-free way (under consistency)
- Biggest unresolved downside is conditional crutch: parameter is conditional on D_1 (first half of data). Important practice issue. Model-free inference for marginal parameter is an open problem

Acknowledgments

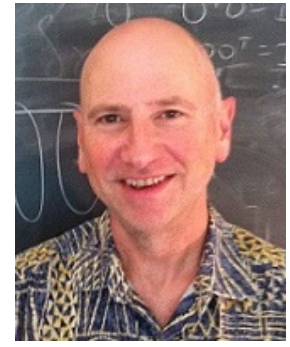
M. G'Sell



J. Lei



A. Rinaldo



L. Wasserman

J. Taylor

R. Tibshirani

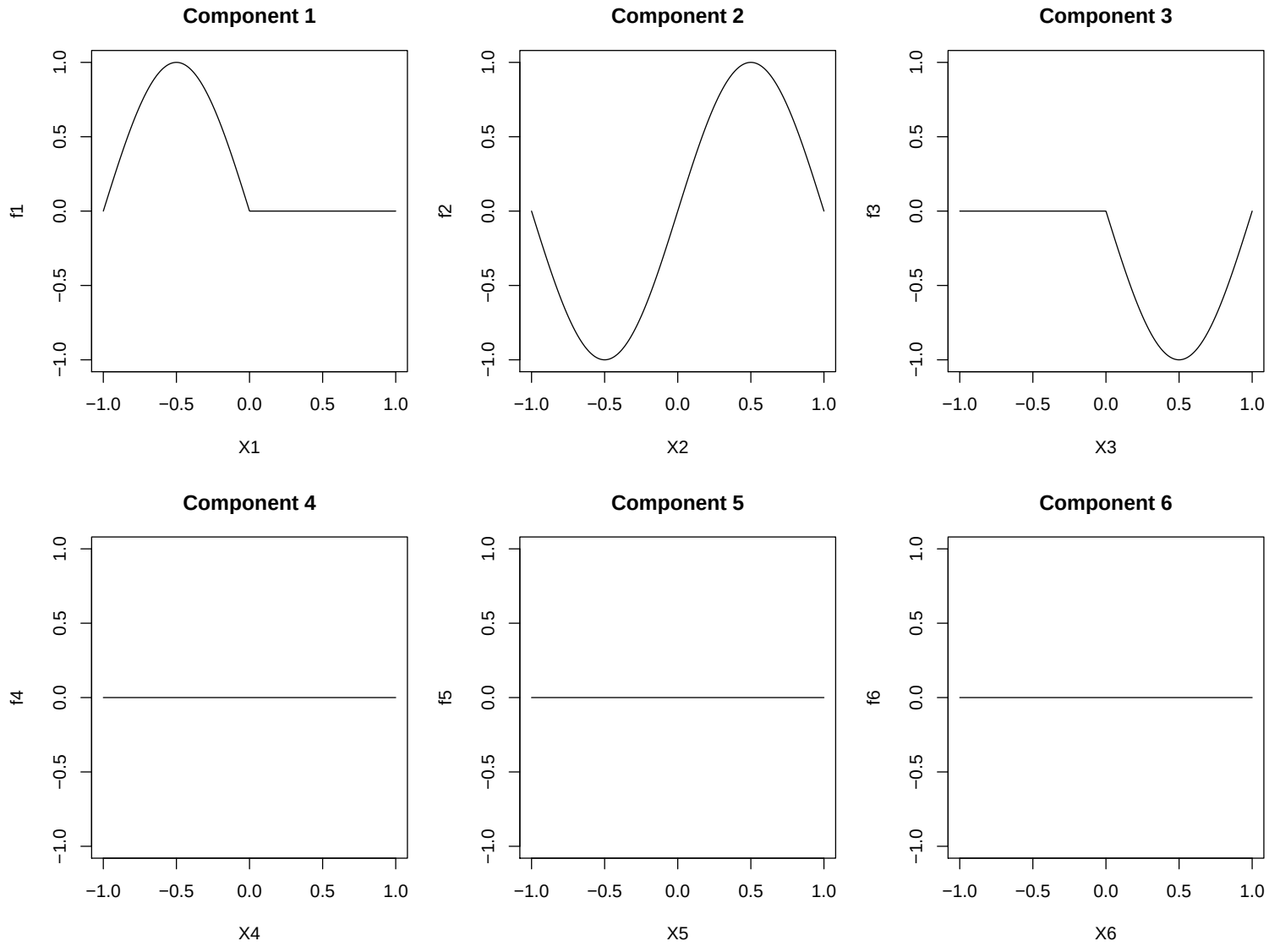
`http://www.stat.cmu.edu/~ryantibs/talks/loco-2018.pdf`

Thank you for listening!

Bonus time

Conformal variable importance

Example: $n = 1000$, $d = 6$, additive model with $f_4 = f_5 = f_6 = 0$



Conformal variable importance

Variable importance intervals $C_j(X_i)$, for $j = 1, \ldots, d$, $i = 1, \ldots, n$