Model-Free Variable Importance

Zhao Lyu, Beining Wu

December 7, 2023

THE UNIVERSITY OF CHICAGO

Leave-One-Covariate-Out (LOCO) [LGR+18]

- Set up:
  - A training data index set: $\mathcal{I}_1 \subseteq \{1, \ldots, n\}$
  - Training data: $\{(X_i, Y_i)\}_{i \in \mathcal{I}_1}$, where $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$
- Predictive Goal: Use the training data to construct $\widehat{\mu}$, our estimate of the mean function $\mu(x) = \mathbb{E}(Y \mid X = x)$

$$\widehat{\mu} = \mathcal{A}(\{(X_i, Y_i)\}_{i \in \mathcal{I}_1})$$

- Selective Goal: Use the training data to select covariates important to the prediction

$$\widehat{\mathcal{I}} = \mathcal{A}'(\{(X_i, Y_i)\}_{i \in \mathcal{I}_1}) \subseteq [d]$$

LOCO-continued

- New data: $\{(X_i(-j), Y_i)\}_{i \in \mathcal{I}_1}$, where $X_i(-j) = (X_i(1), \ldots, X_i(j-1), X_i(j+1), \ldots, X_i(d) \in \mathbb{R}^{d-1}$
- New estimate: $\widehat{\mu}_{(-j)} = \mathcal{A}(\{(X_i(-j), Y_i)\}_{i \in \mathcal{I}_1})$
- Excess prediction error of covariate j at (X_{n+1}, Y_{n+1}), a new i.i.d draw:

$$\Delta_j(X_{n+1}, Y_{n+1}) = |Y_{n+1} - \widehat{\mu}_{(-j)}(X_{n+1})| - |Y_{n+1} - \widehat{\mu}(X_{n+1})|$$

- Interpretation: an increase in prediction error due to not having covariate j in the data set.
- A QUESTION: could $\Delta_j(X_{n+1}, Y_{n+1})$ be negative? How to interpret this?

Local measure of Variable Importance

- Intuition: If covariate j is important, Δ_j should be large
- Method: Construct a valid prediction interval for $\Delta_j(X_{n+1}, Y_{n+1})$:

  $$W_j(x) = \{|y - \widehat{\mu}_{-j}(x)| - |y - \widehat{\mu}(x)| \ : \ y \in C(x)\},$$

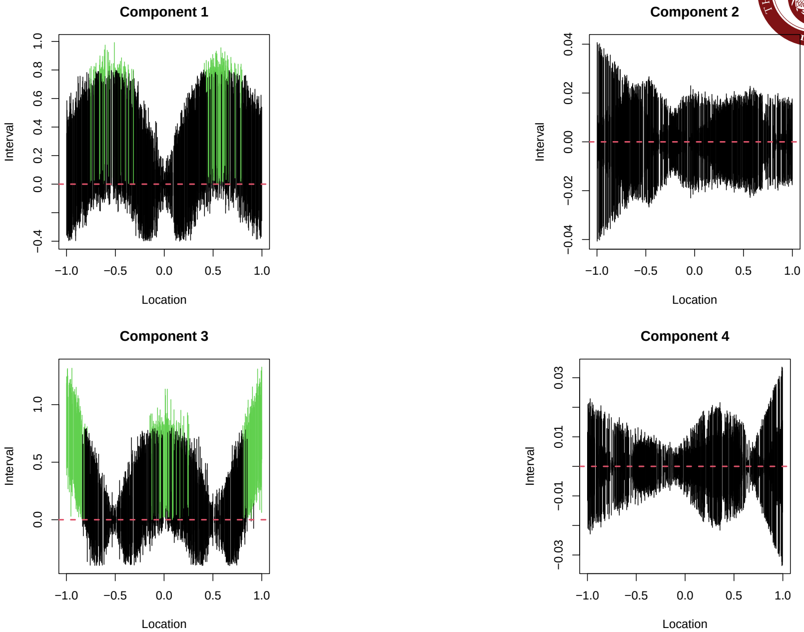
  where C is a conformal prediction set with $\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$.
- Validity: $\forall j \in [d]$,

  $$\mathbb{P}(\Delta_j(X_{n+1}, Y_{n+1}) \in W_j(X_{n+1})) \geq 1 - \alpha.$$

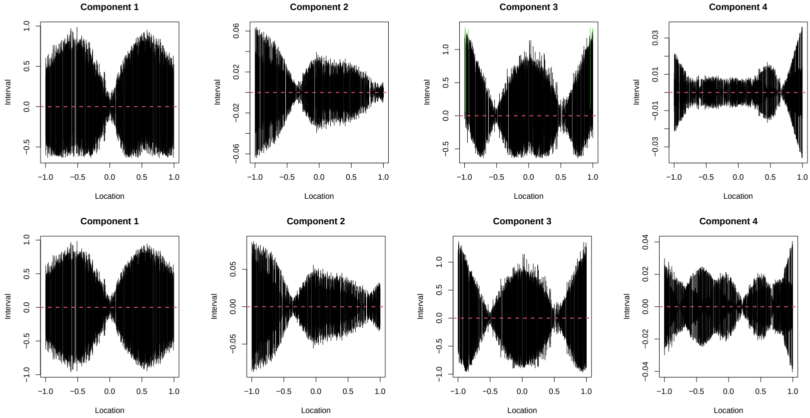
Demo 1: independent X_i in low-dim LOCO via local measure

- Let $X_i \sim \text{Unif}[-1,1]^d$ with $d = 4$.
- Let 1st and 3rd covariates be the important ones by constructing
$$\mu(x) = \sum_{i=1}^{4} f_i(x(i)),$$
where $f_1(t) = \sin(\pi t), f_3(t) = \cos(\pi t)$ and $f_2 = f_4 = 0$.
- Set the response for $i \in [n]$, and set $n = 1000$
$$Y_i = \mu(X_i) + \epsilon_i$$

Demo 1-continued: $\epsilon_i \sim N(0, 0.1)$



Demo 1-continued: if there comes more noise



Demo 2: correlated X_i in low-dim LOCO

- Let $X \sim \mathrm{N}(0, \Sigma)$ with $d = 4$ and $\Sigma = \begin{bmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.
- Let 1st and 3rd covariates be the important ones by constructing

$$\mu(x) = \sum_{i=1}^{4} f_i(x(i)),$$

where $f_1(t) = 100t^2, f_3(t) = 50|t|$ and $f_2 = f_4 = 0$.

- Set the response for $i \in [n]$, and set $n = 1000$

$$Y_i = \mu(X_i) + \epsilon_i, \epsilon_i \sim \mathrm{N}(0, 0.1)$$

Demo 2-continued: changes on ρ

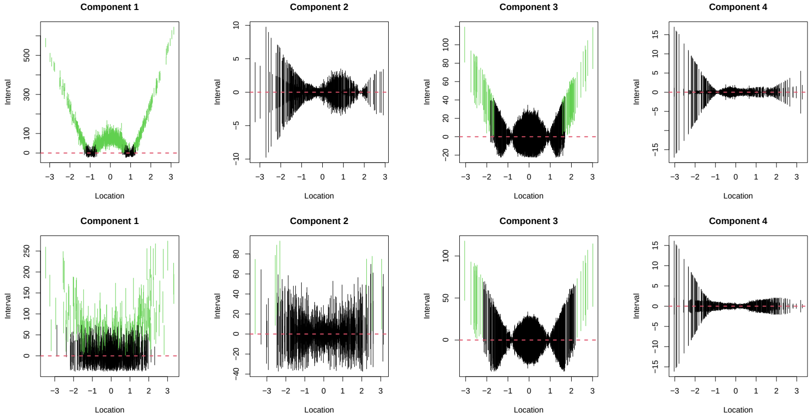


Figure: When $\rho = 0.1$ and $\rho = 0.9$

Conclusion from Demo 1 & 2

- Question: when LOCO work when LOCO does not work?
- (Incomplete) answer:
  - Signal to Noise Ratio(SNR): in Demo 1, when SNR is small (i.e. noise is relatively large), LOCO fails to identify the important covariates: $X_j \not\perp\!\!\!\perp Y \mid X_{-j}$
  - Suprious correlation: in Demo 2, when trivial covariates are strongly correlated with important covariates, LOCO fails: $X_j \perp\!\!\!\perp Y \mid X_{-j}$

Global Measure: Inference for Distribution of Δ_j

- Inference target:

$$\Delta_j(X_{n+1}, Y_{n+1}) = |Y_{n+1} - \widehat{\mu}_{-j}(X_{n+1})| - |Y - \widehat{\mu}(X_{n+1})|.$$

- Previous focus: cover $\Delta_j(X_{n+1}, Y_{n+1})$ with a predictive interval.
- Question: what can we say about the distribution conditioned on the training data, i.e. $\Delta_j(X_{n+1}, Y_{n+1})|\mathcal{D}_0$?
- Local: one-shot coverage - Global: inference for the distribution of $\Delta_j(X_{n+1}, Y_{n+1})|\mathcal{D}_0$.

Global Measure: How?

- Under $H_0 : X_j \perp\!\!\!\perp Y \mid X_{-j}$, we have

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X_{-j}],$$

  so $\Delta_j(X_{n+1}, Y_{n+1})$ (with the perfect knowledge) should be zero. If $\widehat{\mu}$ and $\widehat{\mu}_{-j}$ can reasonably approximate the conditional mean. Then $\theta_j = \mathbb{E}[\Delta_j|\mathcal{D}_0]$ should be hovering around 0.
- Under H_1, from Jensen's inequality, $\theta_j = \mathbb{E}[\Delta_j|\mathcal{D}_0] \geq 0$.
- Heuristic approach:
  1. Train $\widehat{\mu}$ and $\widehat{\mu}_{-j}$ on $\mathcal{D}_0$.
  2. Calculate $\Delta_j(X_i, Y_i), n_0 + 1 \leq i \leq n$.
  3. Form a normal asymptotic confidence interval for θ_j.

Global Measure: How?

- Inference over θ_j requires the existence of first two moments.
- A more robust approach: inference over

$$m_j = \text{median}[\Delta_j(X_{n+1}, Y_{n+1})|\mathcal{D}_0].$$

- Test the hypothesis

$$H_0 : m_j \leq 0 \quad \longleftrightarrow \quad H_1 : m_j > 0,$$

with the Wilcoxon signed-rank test.

Global Measure: Examples

This example is inherited from the original paper [LGR+18]. Consider the data from sparse linear model:

- $X \sim \mathrm{N}(0, I_p)$
- $Y = X^\top \beta + \epsilon$, where $\epsilon \sim \mathrm{N}(0, 1)$ and β is a sparse vector with s non-zero entries.
- We generate $n = 400$ samples with $p = 100$ and $s = 5$. For the coefficient, we set $\beta_j = j/2, j \leq s$ and $\beta_j = 0$ otherwise.
- We perform LOCO procedure with cross-validated Lasso, on the selected variables in the initial Lasso estimate.

Global Measure: Examples

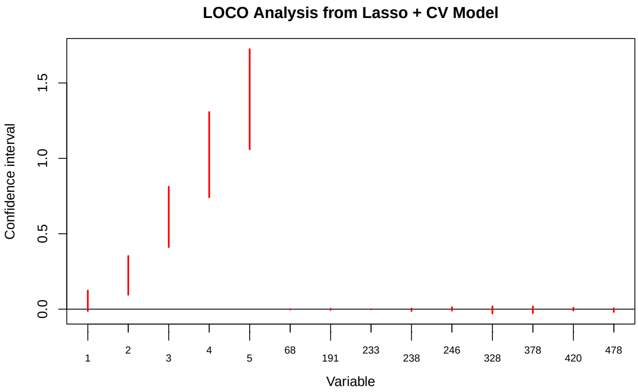


Figure: CI for $\mathrm{med}[\Delta_j|\mathcal{D}_0]$ from Wilcoxon signed-rank test with $\alpha = 0.1$.

Global Measure: False Positive from Correlation and Low SNR

- When there is high design correlation, especially in high dimensional regime, inference procedure always tends to overestimate the importance of a conditionally independent covariate.
- We inspect this issue in the LOCO procedure.
- In the sequel, we set $\beta = 0.25 \times (\mathbf{1}_s^\top, \mathbf{0}_{p-s}^\top)^\top$ and set $X \sim \mathrm{N}(0, (1-\rho)I + \rho \mathbf{1}\mathbf{1}^\top)$

Global Measure: False Positive from Correlation and Low SNR

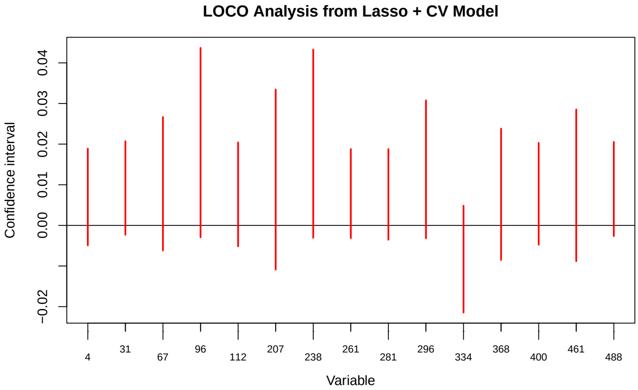


Figure: CI for $\mathrm{med}[\Delta_j|\mathcal{D}_0]$ from Wilcoxon signed-rank test with $\alpha = 0.1$.

Global Measure: False Positive from Correlation and Low SNR

Figure 3 shows that when the correlation is high, one find it hard to distinguish from the important covariates (indexed 1 to 5) and the null covariates (rest) from the LOCO procedure.

Global Measure: False Positive from Model Misspecification

Model Misspecification Causes Systematic Bias

- As we will show below, misspecification is a important source of systematic bias, causing both false positive and false negative.
- To start with, we consider a distribution that could cause the linear model to falsely identify a covariate as important.

Global Measure: False Positive/Negative from Model Misspecification

- Let $X_2 \sim \mathsf{N}(0,1)$, $X_1, Y | X_2 \overset{\text{i.i.d.}}{\sim} \mathsf{N}(X_2^2, 1)$.
- $Y \perp\!\!\!\perp X_1 | X_2$, $Y \not\perp\!\!\!\perp X_2 | X_1$.
- Model class: $\mathcal{F}_{\mathrm{lin}} = \{f : f(x) = x^\top \beta \text{ for some } \beta \in \mathbb{R}^2\}$.
- Linear model is misspecified, but X_1 can interpret the quadratic signal in Y provided by X_2, thus will potentially be identified as important.

Global Measure: False Positive from Model Misspecification

We can look to the bias from the conditional expectation:

$$\mathbb{E}[Y|X] = X_2^2; \quad \mathbb{E}[Y|X_{(-1)}] = X_2^2.$$

Best linear approximation:

$$\begin{aligned}
&\operatorname*{argmin}_{\beta \in \mathbb{R}^2} \mathbb{E}[(\mathbb{E}[Y|X] - \beta_1 X_1 - \beta_2 X_2)^2] \\
&\quad = \operatorname*{argmin}_{\beta} 3(\beta_1 - 1)^2 + \beta_1^2 + \beta_2^2 = (3/4, 0)^\top; \\
&\operatorname*{argmin}_{\beta_2 \in \mathbb{R}} \mathbb{E}[(\mathbb{E}[Y|X_{(-1)}] - \beta_2 X_2)^2] = 0.
\end{aligned}$$

When we restrict ourselves to the linear models, X_1 is important in predicting Y, even if X_1 is independent of Y given X_2!

Global Measure: False Positive from Model Misspecification

We generate $n = 200$ samples from the distribution, and perform LOCO with $\mathcal{F}_{\mathrm{lin}}$, trained by least squares. In the correct model, we include $X_3 = X_2^2$ as the predictor.

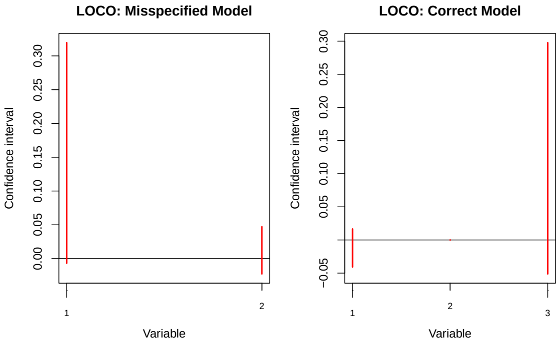


Figure: CI for $\mathrm{med}[\Delta_j|\mathcal{D}_0]$ from Wilcoxon signed-rank test with $\alpha = 0.1$.

Global Measure: False Negative from Model Misspecification

Misspecification from latent variable

- There are some cases where the conditional mean is unable to capture the conditional dependence structure.
- A rather artificial example:
  - Let X_1, X_2 be independent standard normal.
  - $Y = X_1 \cdot \epsilon + X_2 + \tilde{\epsilon}$, where $\epsilon, \tilde{\epsilon} \sim \mathrm{N}(0, 1)$ **are unobserved**.
  - We have

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X_{(-1)}] = X_2$$

  - But $X_1 \not\perp\!\!\!\perp Y \mid X_{(-1)}$.
- In that case, $\Delta_1(X_{n+1}, Y_{n+1})$ cannot capture the conditional dependence, causing **potential false negative.**
- This is different from the previous example, where $X_1 \perp\!\!\!\perp Y \mid X_{(-1)}$

Global Measure: False Negative from Model Misspecification

Would including ϵ as predictor help? Not necessarily as the model class is still misspecified.

Approximation error from a misspecified model class

- Let's continue with the previous example, now we denote X_3 as the ϵ and is observed.
- $\mathbb{E}[Y|X] = X_1 \cdot X_3 + X_2$; $\mathbb{E}[Y|X_{(-1)}] = X_2$. Different!
- Model class: $\mathcal{F}_{\mathrm{lin}} = \{f : f(x) = x^\top \beta \text{ for some } \beta \in \mathbb{R}^3\}$.
- Similar to the case in false positive. There are limits for $\widehat{\mu}$ and $\widehat{\mu}_{(-1)}$ to approximate the conditional mean reasonably.

Global Measure: False Negative from Model Misspecification

Best linear approximator:

$$\operatorname*{argmin}_{f\in\mathcal{F}_{\mathrm{lin}}} \mathbb{E}[(\mathbb{E}[Y|X] - f(X))^2] = X_2$$

$$\operatorname*{argmin}_{f\in\mathcal{F}_{\mathrm{lin}}} \mathbb{E}[(\mathbb{E}[Y|X_{(-1)}] - f(X))^2] = X_2$$

A heuristic calculation indicates that even we include all the important variables for predicting Y, a misspecified model based $\Delta_j|\mathcal{D}_0$ still fails to capture the conditional dependence.

But here, a misspecified model produces the same estimator, and thus produces **false negative** instead of false positive.

Global Measure: False Negative from Model Misspecification

In comparison, we consider the correctly speccified model class:

- Let $\tilde{X} = (X_1, X_2, X_3, X_1X_2, X_2X_3, X_3X_1)$.

$$\mathcal{F} = \{f : f(x) = \tilde{x}^\top \beta \text{ for some } \beta \in \mathbb{R}^6\}.$$

In the seuqel, we generate $n = 200$ samples from the distribution introduced above and perform LOCO procedure with three different specifications introduced above.

Global Measure: False Negative from Model Misspecification

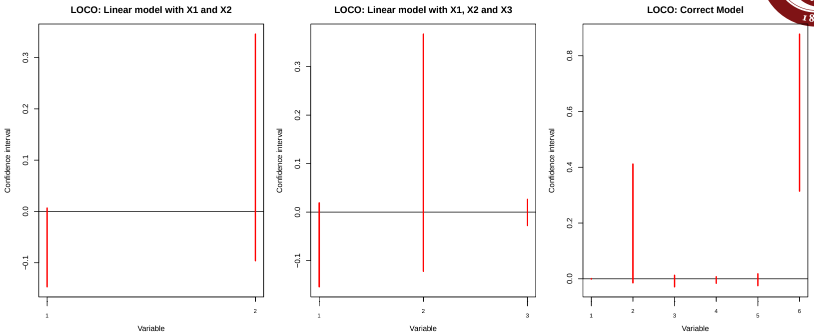


Figure: CI for $\mathrm{med}[\Delta_j | \mathcal{D}_0]$ from Wilcoxon signed-rank test with $\alpha = 0.1$.

In Figure 5, LOCO procedure with only X with linear models fails to identify X_1 and X_3 as significant. But when we include the cross term in the linear model, the procedure produces reasonable results.

Global Measure: False Positive/Negative from Model Misspecification

- The examples above suggested that we should conduct a parallel diagnostic on the models we use for training.
- This non-parametric version of variable significance, or conditional independence itself, cannot be fully measured by one specific quantity.
- This θ_j or m_j metric is only one metric for quantifying the conditional independence. As the previous example shows, sometimes it's not suitable in characterizing the conditional independence.

Reference

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman, *Distribution-free predictive inference for regression*, Journal of the American Statistical Association **113** (2018), no. 523, 1094–1111.