Floodgate: Inference for Model-Free Variable Importance

Bernoulli-IMS One World Symposium 2020

Lu Zhang, Lucas Janson

Department of Statistics, Harvard University

Overview

We introduce floodgate, a new inferential approach for variable importance.

- Focus on an interpretable, sensitive and nonparametric measure of variable importance: the mMSE gap.
- Provide valid and robust lower confidence bounds (LCB) for the mMSE gap.
- Can leverage flexible regression algorithms with good predictive performance to improve inferential accuracy.

Motivation

Setup: data (Y, X, Z) from some joint distribution.

- Y a response variable of interest.
- X a explanatory variable of interest (AKA treatment, covariate, feature).
- $Z := (Z_1, \cdots, Z_p)$ a set of p further variables (AKA confounders, nuisance variables).

Question: Is the variable important or not?

$\Downarrow$

Go beyond: How important is the variable?

Q1: How to define a good measure of variable importance (MOVI)?

Q2: How to provide inference for it?

A desirable MOVI (of the covariate X) should have

Validity: zero when $Y \perp\!\!\!\perp X \mid Z$.

Sensitivity: able to detect nonlinear effects and interactions.

Interpretability: interpretable for scientists and practitioners' use.

A desirable inferential procedure for the MOVI should be:

General Accurate Robust

Our MOVI: the mMSE Gap

The minimum mean squared error (mMSE) gap for variable X is defined as

$$\mathcal{I}^2 = \mathbb{E}\left[(Y - \mathbb{E}[Y \mid Z])^2\right] - \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X, Z])^2\right].$$

We have

$$\mathcal{I}^2 = 0 \iff \mathbb{E}[Y \mid X, Z] \overset{a.s.}{=} \mathbb{E}[Y \mid Z],$$

and the following interpretations:

- **Predictive**: immediate from above.
- **Variance decomposition**: $\mathcal{I}^2 = \mathrm{Var}(\mathbb{E}[Y \mid X, Z]) - \mathrm{Var}(\mathbb{E}[Y \mid Z])$.
- **Causal**: $\mathcal{I}^2 = \frac{1}{2}\mathbb{E}_{x_1, x_2 \overset{i.i.d.}{\sim} P_{X|Z}}\left[(\mathbb{E}[Y \mid X = x_1, Z] - \mathbb{E}[Y \mid X = x_2, Z])^2\right]$.
- Compact form: $\mathcal{I}^2 = \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid X, Z] \mid Z)]$.

Main Methodology: Floodgate

True regression function $\mu^\star(x, z) := \mathbb{E}[Y \mid X = x, Z = z]$

$$\Rightarrow \quad \mathcal{I}^2 = \mathbb{E}[\mathrm{Var}(\mu^\star(X, Z) \mid Z)] = \mathbb{E}\left[(\mu^\star(X, Z) - \mathbb{E}[\mu^\star(X, Z) \mid Z])^2\right]$$

Challenges:

- $\mu^\star$ unknown.
- Nonlinearity in the above functional.

Possible solution: ~~assume we have a good estimator μ of $\mu^\star$~~?

The idea of floodgate:

(a) Construct a functional f such that $f(\mu) \le \mathcal{I}$ for any μ.

(b) Know how to obtain LCB $L(\mu)$ of $f(\mu)$ for any μ.

(c) (Ideally) the functional f also satisfies $f(\mu^\star) = \mathcal{I}$.

Our choice of floodgate functional (to satisfy (a) and (c)):

$$f(\mu) := \frac{\mathbb{E}[\mathrm{Cov}(\mu^\star(X, Z), \mu(X, Z) \mid Z)]}{\sqrt{\mathbb{E}[\mathrm{Var}(\mu(X, Z) \mid Z)]}}.$$

Our assumption (to make (b) possible): $P_{X|Z}$ known (note we also have robustness analysis and assumption relaxation).

Lemma (A deterministic relationship)

For any μ such that $f(\mu)$ exists, $f(\mu) \le \mathcal{I}$ and $f(\mu^\star) = \mathcal{I}$.

Algorithm 1 Floodgate

Input: $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, $P_{X|Z}$, μ, confidence level $\alpha \in (0, 1)$.

Compute, for each $i \in [n]$, $R_i = Y_i\big(\mu(X_i, Z_i) - \mathbb{E}[\mu(X_i, Z_i) \mid Z_i]\big)$, $V_i = \mathrm{Var}(\mu(X_i, Z_i) \mid Z_i)$ and their sample mean $(\bar{R}, \bar{V})$ and sample covariance matrix $\hat{\Sigma}$, and compute $s^2 = \frac{1}{\bar{V}}\left[\left(\frac{\bar{R}}{2\bar{V}}\right)^2 \hat{\Sigma}_{22} + \hat{\Sigma}_{11} - \frac{\bar{R}}{\bar{V}}\hat{\Sigma}_{12}\right]$.

Output: Lower confidence bound $L_n^\alpha(\mu) = \max\left\{\frac{\bar{R}}{\sqrt{\bar{V}}} - \frac{z_\alpha s}{\sqrt{n}}, 0\right\}$, with the convention that $0/0 = 0$.

More computation details:

- μ can be fitted from a separate dataset e.g. via sample splitting.
- Generally, draw $\tilde{X}^{(k)}, k = 1, \cdots, K$ from $P_{X|Z}$, conditionally independently of X, Y then plug-in the Monte Carlo estimators.

Theorem (Asymptotic validity)

Under mild moment conditions on Y and $\mu(X, Z)$, we have

$$\mathbb{P}\left(L_n^\alpha(\mu) \le \mathcal{I}\right) \ge 1 - \alpha - O(n^{-1/2}).$$

Accuracy: inferential accuracy is directly related to the MSE of "μ_n".

Floodgate procedure is invariant respect to a "equivalent" function class of μ, $S_\mu = \{c\mu(x, z) + g(z) : c > 0, g : \mathbb{R}^p \to \mathbb{R}\}$.

Under mild moment conditions on Y and noises, for μ_n with well-behaved moments,

$$\mathcal{I} - L_n^\alpha(\mu_n) = O_p\left(\inf_{\mu \in S_{\mu_n}} \mathbb{E}\left[(\mu(X, Z) - \mu^\star(X, Z))^2\right] + n^{-1/2}\right).$$

Main Methodology: Floodgate

Robustness: floodgate is robust to the estimation error of $P_{X|Z}$.

Suppose $P_{X|Z}$ unknown, we instead use its estimate $Q_{X|Z}^{(n)}$.

Under moment conditions on Y and noises, for μ_n with well-behaved moments under both the true distribution and the specified one, we have

$$\mathbb{P}\left(L_n^\alpha(\mu_n) \le \mathcal{I} + \Delta_n\right) \ge 1 - \alpha - O(n^{-1/2}),$$

where

$$\Delta_n \le c_1\sqrt{\mathbb{E}\left[\chi^2\left(P_{X|Z} \,||\, Q_{X|Z}^{(n)}\right)\right]} - c_2\mathbb{E}\left[(\bar{\mu}_n(X, Z) - \mu^\star(X, Z))^2\right]$$

where $\bar{\mu}_n$ is a particular representative of S_{μ_n} and $\chi^2(\cdot \,||\, \cdot)$ denotes the χ^2 divergence.

Application to Genomic Study of Platelet Count

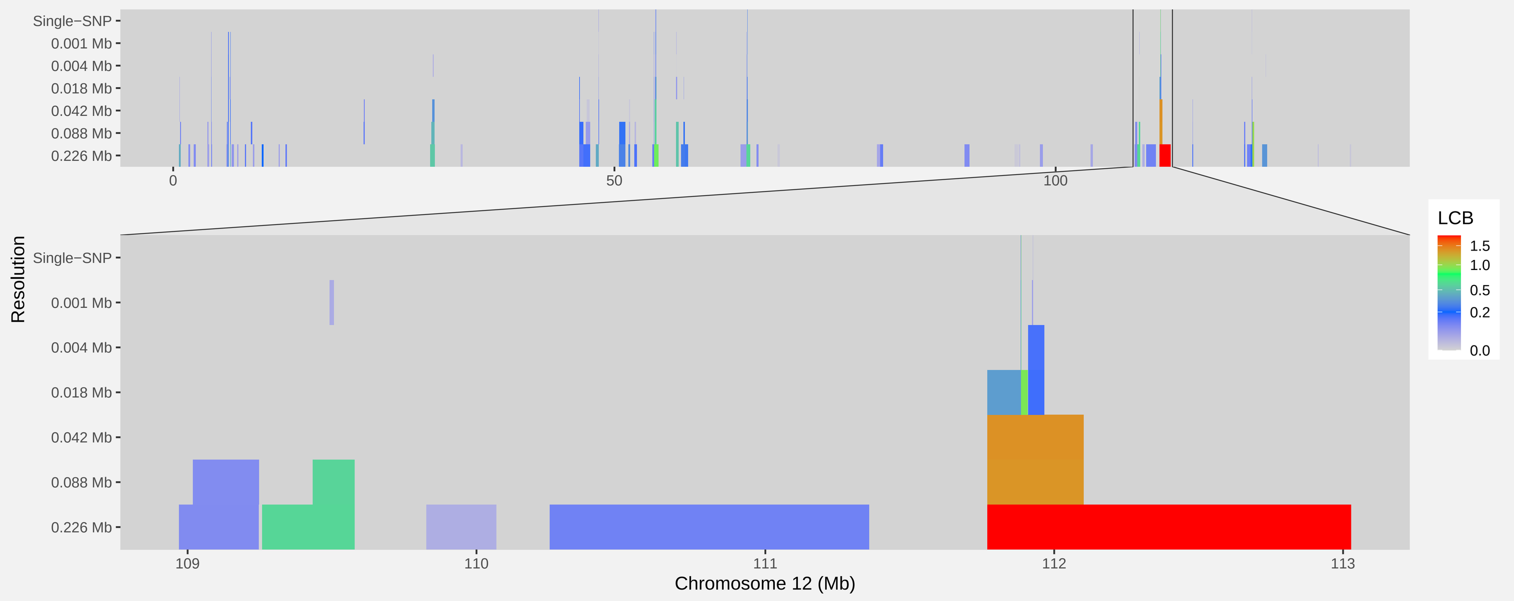

Figure 1:Colored Chicago plot [1] with the color of each point representing the floodgate LCB for the importance of a group of SNPs on Chromosome 12 at different resolutions (y-axis). Bottom plot shows a zoomed-in region of strong importance.

Extensions

1. Co-sufficient floodgate relaxes the assumptions to only knowing a model for $P_{X|Z}$
2. Floodgate for a different measure of variable importance.
3. Inference on group variable importance.
4. Transporting floodgate inference to a different covariate distribution.
5. Adjusting for multiplicity and selection effects.

References

[1] Matteo Sesia, Eugene Katsevich, Stephen Bates, Emmanuel Candès, and Chiara Sabatti. Multi-resolution localization of causal variants across the genome. Nature communications, 11(1):1–10, 2020.

[2] Lu Zhang and Lucas Janson. Floodgate: Inference for model-free variable importance. arXiv preprint arXiv:2007.01283, 2020.

Acknowledgement: L.Z. is partially supported by the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard, award number #1764269 and the Harvard Quantitative Biology Initiative. L.J. is partially supported by the William F. Milton Fund.