Nonparametric Interaction Search, Quick and Easy

Feng Ruan

(Joint work with Keli Liu)

Keli Liu

Missing Heritability Problem

The mystery of missing heritability: Genetic interactions create phantom heritability

Or Zuk[a], Eliana Hechter[a], Shamil R. Sunyaev[a,b], and Eric S. Lander[a,1]

Broad Institute of MIT and Harvard, Cambridge, MA 02142; and [b]Genetics Division, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115

Contributed by Eric S. Lander, December 5, 2011 (sent for review October 9, 2011)

Human genetics has been haunted by the mystery of "missing heritability" of common traits. Although studies have discovered >1,200 variants associated with common diseases and traits, these variants typically appear to explain only a minority of the heritability. The proportion of heritability explained by a set of variants is the ratio

(frequency <1%) with large effects (3–9). We will discuss the frequency spectrum of disease-related variants in our second paper in this series.

Here we explore the possibility that a significant portion of the missing heritability might not reflect missing variants at all. The

- 80% of the currently missing heritability for Crohn's disease could be due to genetic *interactions*.

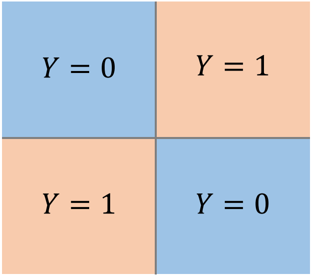
What are Interactions?

Interaction: Effect of one variable depends on the other variables

Example: XOR Signal

$$Y = \begin{cases} 1 & \text{if } X_1 X_2 > 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{P}(X_j = \pm 1) = 1/2$$



- $X_1 \perp Y$ and $X_2 \perp Y$ but $(X_1, X_2) \not\perp Y$.

The Linear Model View of Interactions

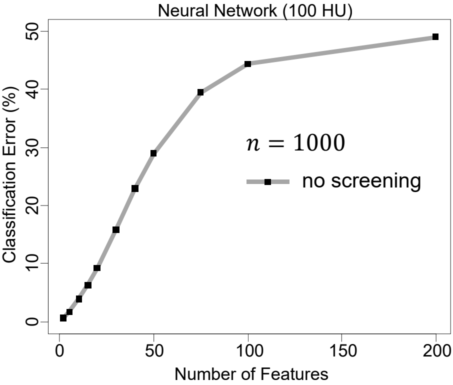
$$\text{logit}\,\mathbb{P}(Y=1|X) = \alpha + \sum_j \gamma_j \cdot X_j + \sum_{j<k} \theta_{jk} \cdot X_j \cdot X_k$$

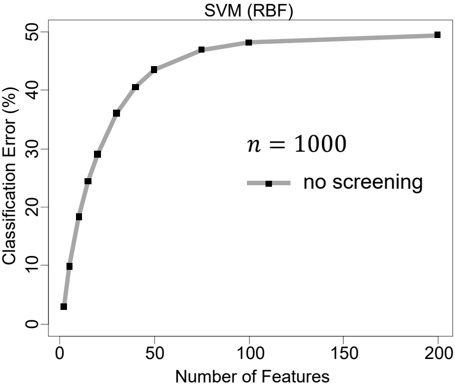
- Problem of Enumeration: $O(p^2)$ terms for order 2 interaction.
- Problem of Specification: Why products $X_j \cdot X_k$? Why not $g(X_j, X_k)$?

Prior Art on Detecting Interactions

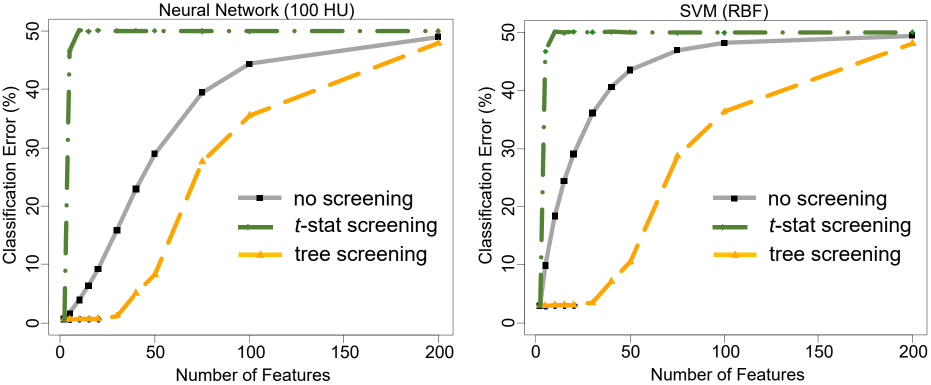
- Linear model:
  - (1) (often) assumes a specific form on the signal and (2) high computation cost.
  - The computation cost can be significantly reduced with special assumptions on the signal/data/model. [Wu et al. 09'; Wu et al. 10'; Shah and Meinshausen 14'; Hao and Zhang 17'; Thanei et al. 18']
- Tree method:
  - (1) nonparametric and (2) low computation cost (linear in p). [Loh 02'; Strobl et al. 08'; Basu et al. 18']
  - Implicit hierarchical assumption on the interaction signals.
- Nonparametric dependence measure:
  - Mutual information: [Póczos and Schneider 12'; Runge 18']
  - Kernel based measure: [Gretton et al. 07', Gretton et al. 08', Fukumizu et al. 09'].
  - Distance correlation: [Székely and Rizzo 14', Fan et al. 15']
  - Others: [Azadkia and Chatterjee 19'].
- Neural network (NN).

Order 2 XOR: NNets/SVMs Don't Work in High Dim.

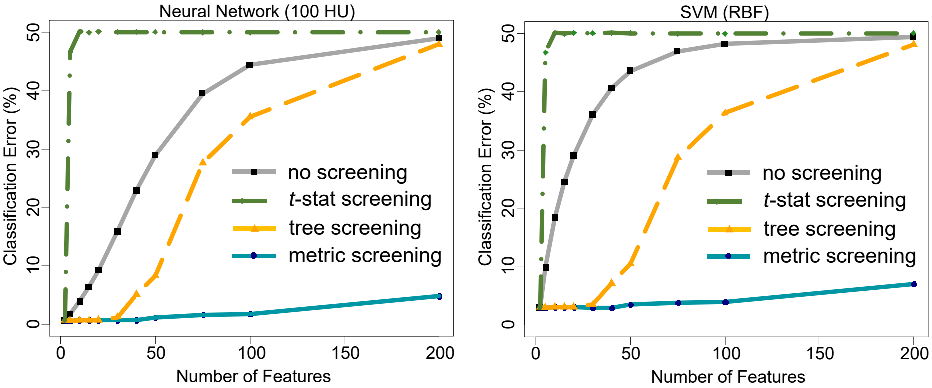




Order 2 XOR: Trees Assume Hierarchical Signals



The Algorithm of this Talk: Metric Screening



The Goals

Goal 1: Nonparametric—agnostic to (complex) form of signals.

Goal 2: Reasonable power to detect signals.

Goal 3: Computation cost is linear in p.

THIS TALK: let's achieve these goals (when Y is binary).

Part I

Metric Learning Algorithm

Idea 1: Nonparametric Two Sample Test

Goal: We want to select signal variables X_S out of all variables X.

Key Observation: Difference between signal X_S and noise variables X_{S^c}.

- As signal, X_S satisfy $\mathcal{L}(X_S \mid Y=1) \neq \mathcal{L}(X_S \mid Y=0)$.
- As noise, X_{S^c} satisfy $\mathcal{L}(X_{S^c} \mid Y=1) = \mathcal{L}(X_{S^c} \mid Y=0)$.

Starting Point: Problem of selecting signals $\Leftrightarrow$ Problem of two sample test

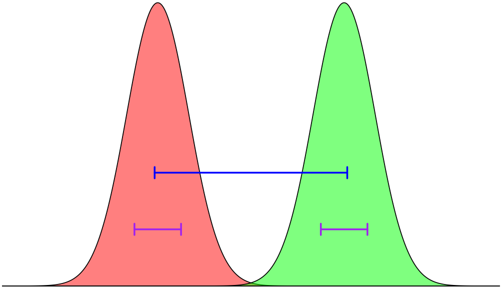
$$H_0 : \mathbb{P}_1 = \mathbb{P}_0 \text{ vs. } H_1 : \mathbb{P}_1 \neq \mathbb{P}_0.$$

Nonparametric Two Sample Test: $\mathbb{P}_1 = \mathbb{P}_0$ vs. $\mathbb{P}_1 \neq \mathbb{P}_0$

Distance covariance test (Székely and Rizzo 05): compute the measure

$$\mathbb{E}_{B-W}[\|X - X'\|_2]$$

$$:= \underbrace{\mathbb{E}[\|X - X'\|_2 \mid Y \neq Y']}_{\mathbb{E}_B[\|X-X'\|_2]} - \underbrace{\mathbb{E}[\|X - X'\|_2 \mid Y = Y']}_{\mathbb{E}_W[\|X-X'\|_2]}.$$



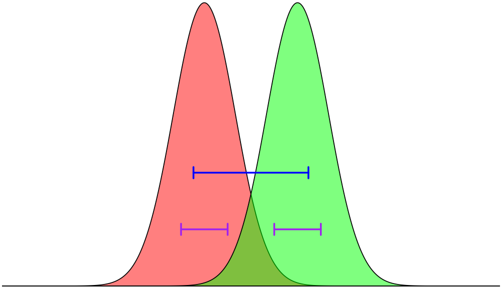
$$\mathbb{E}_B[\|X - X'\|_2] \gg \mathbb{E}_W[\|X - X'\|_2]$$

Nonparametric Two Sample Test: $\mathbb{P}_1 = \mathbb{P}_0$ vs. $\mathbb{P}_1 \neq \mathbb{P}_0$

Distance covariance test (Székely and Rizzo 05): compute the measure

$$\mathbb{E}_{B-W}[\|X - X'\|_2]$$

$$:= \underbrace{\mathbb{E}[\|X - X'\|_2 \mid Y \neq Y']}_{\mathbb{E}_B[\|X-X'\|_2]} - \underbrace{\mathbb{E}[\|X - X'\|_2 \mid Y = Y']}_{\mathbb{E}_W[\|X-X'\|_2]}.$$



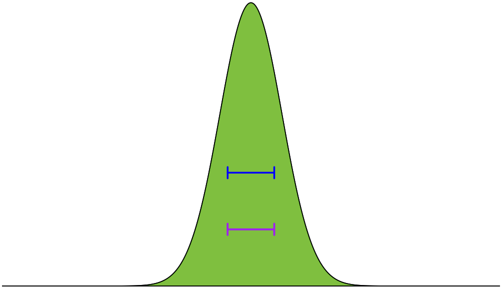
$$\mathbb{E}_B[\|X - X'\|_2] > \mathbb{E}_W[\|X - X'\|_2]$$

Nonparametric Two Sample Test: $\mathbb{P}_1 = \mathbb{P}_0$ vs. $\mathbb{P}_1 \neq \mathbb{P}_0$

Distance covariance test (Székely and Rizzo 05): compute the measure

$$\mathbb{E}_{B-W}[\|X - X'\|_2]$$

$$:= \underbrace{\mathbb{E}[\|X - X'\|_2 \mid Y \neq Y']}_{\mathbb{E}_B[\|X-X'\|_2]} - \underbrace{\mathbb{E}[\|X - X'\|_2 \mid Y = Y']}_{\mathbb{E}_W[\|X-X'\|_2]}.$$



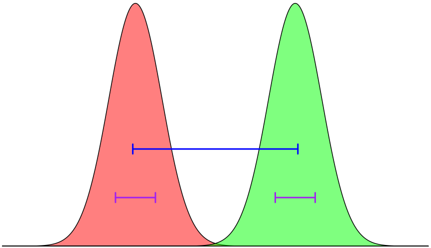
$$\mathbb{E}_B[\|X - X'\|_2] = \mathbb{E}_W[\|X - X'\|_2]$$

Nonparametric Two Sample Test: $\mathbb{P}_1 = \mathbb{P}_0$ vs. $\mathbb{P}_1 \neq \mathbb{P}_0$

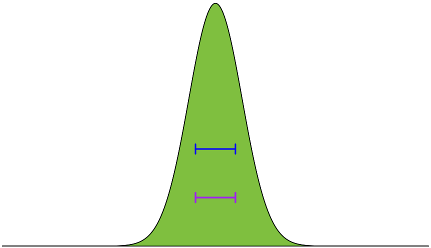
Distance covariance test (Székely and Rizzo 05): compute the measure

$$\mathbb{E}_{B-W}[\|X - X'\|_2]$$

$$:= \underbrace{\mathbb{E}[\|X - X'\|_2 \mid Y \neq Y']}_{\mathbb{E}_B[\|X-X'\|_2]} - \underbrace{\mathbb{E}[\|X - X'\|_2 \mid Y = Y']}_{\mathbb{E}_W[\|X-X'\|_2]}.$$



$$\mathbb{E}_{B-W}[\|X - X'\|_2] \gg 0$$



$$\mathbb{E}_{B-W}[\|X - X'\|_2] = 0$$

Nonparametric Two Sample Test (General Form)

Nonparametric two sample test (general form): compute dependence measure

$$D(\mathbb{P}_0, \mathbb{P}_1) = \mathbb{E}_{B-W}\left[f\big(\|X - X'\|_q^q \big) \right].$$

Example

- Distance covariance test: $f(x) = \sqrt{x}$ and $q = 2$

$$\mathbb{E}_{B-W}\left[\|X - X'\|_2 \right].$$

- MMD test with RBF kernel: $f(x) = -\exp(-x)$ and $q = 2$

$$-\mathbb{E}_{B-W}\left[\exp(-\|X - X'\|_2^2) \right].$$

Nonparametric Two Sample Test (General Form)

Nonparametric two sample test (general form): compute dependence measure

$$D(\mathbb{P}_0, \mathbb{P}_1) = \mathbb{E}_{B-W}\left[f\left(\|X - X'\|_q^q \right) \right].$$

Theorem (Liu and R. 20')

Let $q \in \{1, 2\}$. Then $D(\mathbb{P}_0, \mathbb{P}_1)$ is a valid dependence measure if and only if

f' is strictly complete monotone, i.e., $(-1)^{k-1} f^{(k)}(x) > 0$ for all $k \geq 1$.

Definition (Valid Dependence Measure)

1. $D(\mathbb{P}_0, \mathbb{P}_1) \geq 0$ for all $\mathbb{P}_0, \mathbb{P}_1$.
2. $D(\mathbb{P}_0, \mathbb{P}_1) = 0$ if and only if $\mathbb{P}_0 = \mathbb{P}_1$.

Proof: The proof follows classical arguments by Bernstein and Schoenberg.

Back to Feature Selection

Goal: We want to select signal variables X_S out of all variables X.

Key: $\mathcal{L}(X_S \mid Y = 1) \neq \mathcal{L}(X_S \mid Y = 0)$.

Back to Feature Selection

Goal: We want to select signal variables X_S out of all variables X.

$$\text{Key: } \mathcal{L}(X_S \mid Y = 1) \neq \mathcal{L}(X_S \mid Y = 0).$$

A First Idea (Song et al. 12'):

Find the subset $T \subseteq \{1, 2, \ldots, p\}$ that maximizes the dependence measure:

$$\max_T \; \mathbb{E}_{B-W}\left[f(\|X_T - X_T'\|_q^q) \right].$$

Back to Feature Selection

Goal: We want to select signal variables X_S out of all variables X.

$$\text{Key: } \mathcal{L}(X_S \mid Y = 1) \neq \mathcal{L}(X_S \mid Y = 0).$$

A First Idea (Song et al. 12'):

Find the subset $T \subseteq \{1, 2, \ldots, p\}$ that maximizes the dependence measure:

$$\max_T \ \mathbb{E}_{B-W}\left[f(\|X_T - X'_T\|_q^q)\right].$$

A Second Idea (this talk):

Find the support of β that maximizes the parameterized dependence measure:

$$\max_\beta \ \mathbb{E}_{B-W}\left[f(\|X - X'\|_{q,\beta}^q)\right]$$

$$\text{where } \ \|X - X'\|_{q,\beta}^q = \sum_j \beta_j |X_j - X'_j|^q.$$

Back to Feature Selection

Goal: We want to select signal variables X_S out of all variables X.

$$\text{Key: } \mathcal{L}(X_S \mid Y=1) \neq \mathcal{L}(X_S \mid Y=0).$$

A First Idea (Song et al. 12'):

Find the subset $T \subseteq \{1, 2, \ldots, p\}$ that maximizes the dependence measure:

$$\max_T \ \mathbb{E}_{B-W}\left[f(\|X_T - X'_T\|_q^q)\right].$$

A Second Idea (this talk):

Find the support of β that maximizes the parameterized dependence measure:

$$\max_\beta \ \mathbb{E}_{B-W}\left[f(\|X - X'\|_{q,\beta}^q)\right]$$

$$\text{where} \quad \|X - X'\|_{q,\beta}^q = \sum_j \beta_j |X_j - X'_j|^q.$$

Good ideas?

Perhaps a Big Surprise

Maximize Dependence Measure

$$\max_{T} \ \mathbb{E}_{B-W}\left[f(\|X_T - X'_T\|_q^q)\right] \ \text{ or } \ \max_{\beta} \ \mathbb{E}_{B-W}\left[f(\|X - X'\|_{q,\beta}^q)\right].$$

Inconsistency Result (Liu and R. 20')

Assume

- You have the power to find the global maximizer.
- You may choose whatever f, q you want.

There exists a distribution $\mathbb{P}$ such that if $(X, Y) \sim P$ then the solution $\hat{S}$ can't find all the signal variables, i.e.,

$$\mathcal{L}(Y|X) \neq \mathcal{L}(Y|X_{\hat{S}}).$$

Perhaps a Big Surprise

Maximize Dependence Measure

$$\max_{T} \ \mathbb{E}_{B-W}\left[f(\|X_T - X'_T\|_q^q)\right] \ \text{ or } \ \max_{\beta} \ \mathbb{E}_{B-W}\left[f(\|X - X'\|_{q,\beta}^q)\right].$$

Inconsistency Result (Liu and R. 20')

Assume

- You have the power to find the global maximizer.
- You may choose whatever f, q you want.

There exists a distribution $\mathbb{P}$ such that if $(X, Y) \sim P$ then the solution $\hat{S}$ can't find all the signal variables, i.e.,

$$\mathcal{L}(Y|X) \neq \mathcal{L}(Y|X_{\hat{S}}).$$

Simply maximizing dependence measure is WRONG!

How to fix it? First Identify the Problem.

Focus on the continuous version:

$$\max_{\beta} \; F(\beta) = \mathbb{E}_{B-W}\left[f(\|X - X'\|_{q,\beta}^{q})\right].$$

$$\hat{S} = \mathrm{supp}(\hat{\beta}) \;\text{ where }\; \hat{\beta} = \mathrm{argmax}\, F(\beta).$$

The Masking Phenomenon

- No false positives: $\hat{S} \subseteq S$.
- May have false negatives (miss signal variables): $Y \mid X_{\hat{S}} \neq Y \mid X_S$.

How to fix it? First Identify the Problem.

Focus on the continuous version:

$$\max_{\beta} \; F(\beta) = \mathbb{E}_{B-W}\left[f(\|X - X'\|_{q,\beta}^{q}) \right].$$

$$\hat{S} = \mathrm{supp}(\hat{\beta}) \;\text{ where }\; \hat{\beta} = \mathrm{argmax}\, F(\beta).$$

The Masking Phenomenon

- No false positives: $\hat{S} \subseteq S$.
- May have false negatives (miss signal variables): $Y \mid X_{\hat{S}} \neq Y \mid X_S$.
- Summary: There is competition between variables.

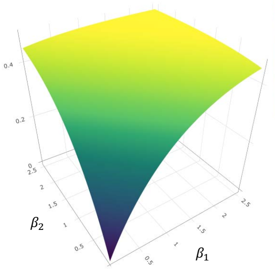
A strong signal will mask weaker signals.

Example: the Landscape of the Objective

$$F(\beta) = \mathbb{E}_{B-W}\left[f(\|X - X'\|_{q,\beta}^{q})\right].$$

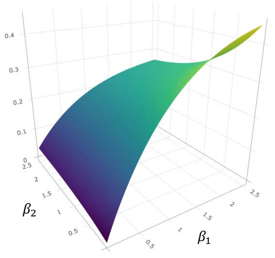
Example

Two signal variables X_1, X_2. The signal is additive across X_1, X_2.



Equal main effects

$\hat{\beta}_1 > 0, \hat{\beta}_2 > 0.$



Effect of X_1 dominates X_2

$\hat{\beta}_1 > 0, \hat{\beta}_2 = 0.$

Idea 2: Reweighting—a solution to the masking effect

$$\mathbb{P} \mapsto \tilde{\mathbb{P}}: \quad \tilde{\mathbb{P}}(x,y) \propto \mathbb{P}(x,y) \cdot w(x,y).$$

$$w(x,y) = \mathbb{P}(Y = 1 - y | X_{\hat{S}} = x_{\hat{S}}).$$

Properties of Reweighting (Liu and R. 20')

- $Y \perp X_{\hat{S}}$ under $\tilde{\mathbb{P}}$.
- $X_{\hat{S}^c} | Y, X_{\hat{S}}$ is the same under $\tilde{\mathbb{P}}$ as under $\mathbb{P}$.

Statistical Implication

- It removes the effect of the selected variables.
- It does not affect the remaining signal.

Iterative Reweighting: $F(\beta; w) = \widetilde{\mathbb{E}}_{B-W}\left[f\left(\|X - X'\|_{q,\beta}^{q}\right)\right]$ $\quad \widetilde{\mathbb{E}}_{B-W}$ w.r.t. $\tilde{\mathbb{P}}$

Metric Screening Algorithm (Population $n = \infty$)

Initialization: $\hat{S} = \emptyset$ and $w_i \equiv 1$.

Iteratively do the following two steps:

1. Maximize dependence measure:

$$\max_{\beta \geq 0} \ F(\beta; w).$$

   Update $\hat{S} \leftarrow \hat{S} \cup \mathrm{supp}(\hat{\beta})$. Stop if $\hat{\beta} = 0$.

2. Reweight the samples: $w_i = 1 - \mathbb{P}(Y = y_i \mid X_{\hat{S}} = x_{i,\hat{S}})$.

Metric Screening Algorithm (Empirical $n < \infty$)

Initialize $\hat{S} = \emptyset$. Initialize the weight $w_i \equiv 1$.

1. Maximize dependence measure:

$$\max_{\beta \geq 0} \; F_n(\beta; w) - \lambda \|\beta\|_1.$$

   Update $\hat{S} \leftarrow \hat{S} \cup \mathrm{supp}(\hat{\beta})$. Stop if $\hat{\beta} = 0$.

2. Reweight the samples: $w_i = 1 - \mathbb{P}(Y = y_i \mid X_{\hat{S}} = x_{i,\hat{S}})$.

Important Remarks (useful for later theoretical discussions):

1. In step 1, we assume it only returns a stationary point since F is nonconvex.
2. In step 2, we assume access to $\mathbb{P}(Y \mid X_A)$ for any subset $A \subseteq S$.

Part II

Theoretical Analysis

Definition: Signal and Noise Variables

Definition (Liu and R. 20')

Let S be the minimal subset such that

- $Y|X = Y|X_S \quad$ (X_S contains all information of X)
- $X_S \perp X_{S^c} \qquad$ (X_{S^c} are irrelevant)

Call X_S the signal variables, and X_{S^c} the noise variables.

Example

Assume the model:

$$Y = g(X_1) = h(X_2), \quad (X_1, X_2) \perp (X_3, \ldots, X_p).$$

Then $X_{\{1,2\}}$ are signal, and $X_{\{3,4,\ldots,p\}}$ are noise variables.

Consistency of Metric Screening

Theorem [Liu and R.' 20]

The metric screening algorithm (on population) returns $\hat{S}$ that satisfies

- No false positive: $\hat{S} \subseteq S$.
- Signal recovery: $\mathcal{L}(Y|X) = \mathcal{L}(Y|X_{\hat{S}})$.

Example

Assume the model:

$$Y = g(X_1) = h(X_2) \quad (X_1, X_2) \perp (X_3, \ldots, X_p)$$

Then $S = \{1, 2\}$, and $\hat{S}$ can be $\{1\}$, $\{2\}$ or $\{1, 2\}$.

From Population $n = \infty$ to Empirical $n < \infty$.

Assumptions:

- X_j is σ_X-subgaussian for $1 \leq j \leq p$.
- Imperfect classification: for some $\rho > 0$

$$\mathbb{E}[\mathbb{P}(Y = 1 - y \mid X_S) \mid Y = y] > \rho \text{ for } y \in \{0, 1\}.$$

- High dimensional regime: $|S| \lesssim \frac{n}{\log p}$.

From Population $n = \infty$ to Empirical $n < \infty$.

Assumptions:

- X_j is σ_X-subgaussian for $1 \leq j \leq p$.
- Imperfect classification: for some $\rho > 0$

$$\mathbb{E}[\mathbb{P}(Y = 1 - y \mid X_S) \mid Y = y] > \rho \text{ for } y \in \{0, 1\}.$$

- High dimensional regime: $|S| \lesssim \frac{n}{\log p}$.

These assumptions are enough to guarantee the *concentration* results.

All Stationary Points Exclude Noise

Theorem (Liu and R. 20')

Consider the metric learning objective: $(w(x,y) \propto \mathbb{P}(Y = 1-y \mid X_A))$

$$\max_{\beta \geq 0} F_n(\beta; w) = \hat{\mathbb{E}}^w_{B-W}\left[f\left(\|X - X'\|^q_{q,\beta}\right)\right] - \lambda \|\beta\|_1$$

Any stationary point β *satisfies* $\mathrm{supp}(\beta) \subseteq S$ *w.h.p, if* $\lambda = \Omega\left(\sqrt{\frac{\log p}{n}}\right)$.

Proof Sketch

- The results holds on population ($n = \infty$).

$$\frac{\partial}{\partial \beta_j} F_\infty(\beta) < 0 \ \text{ for } j \in S^c. \qquad (*)$$

  Any stationary point β of $F_\infty(\beta)$ must have $\mathrm{supp}(\beta) \subseteq S$.

- Uniform convergence transfers the result to finite samples ($n < \infty$).

All Stationary Points Exclude Noise

Theorem (Liu and R. 20')

Consider the metric learning objective: $(w(x, y) \propto \mathbb{P}(Y = 1 - y \mid X_A))$

$$\max_{\beta \geq 0} F_n(\beta; w) = \hat{\mathbb{E}}^w_{B-W}\left[f\left(\|X - X'\|^q_{q,\beta}\right)\right] - \lambda \|\beta\|_1$$

Any stationary point β *satisfies* $\mathrm{supp}(\beta) \subseteq S$ *w.h.p, if* $\lambda = \Omega\left(\sqrt{\frac{\log p}{n}}\right)$.

Proof:

$$\frac{\partial}{\partial \beta_j} F_\infty(\beta) < 0 \text{ for } j \in S^c. \qquad (*)$$

Note: f is completely monotone, i.e., $(-1)^{k-1} f^{(k)}(x) > 0 \Rightarrow$ so is $-f'$.

$$\frac{\partial}{\partial \beta_j} F_\infty(\beta) = \mathbb{E}_{B-W}\left[f'(\|X - X'\|^q_{q,\beta}) \cdot |X_j - X'_j|^q\right] - \lambda$$

$$\overset{j \notin S}{=} \mathbb{E}\Big[\underbrace{\mathbb{E}_{B-W}\left[f'(\|X - X'\|^q_{q,\beta}) \mid X_{S^c}, X'_{S^c}\right]}_{\leq 0} \cdot |X_j - X'_j|^q\Big] - \lambda < 0.$$

Statistical Implications

Consequences:

- Metric learning has no false positive: $\hat{S} \subseteq S$ with high probability!
- If we don't converge to $\beta = 0$, we'll have found true variables!

Remaining Questions:

- Can we find non-zero stationary points when there are true variables?
- Can we design the (non-convex) objective (landscape) so that it is easier for gradient ascent to find non-zero stationary points (true variables)?

Idea 3: Design the Landscape of the Objective

The objective:

$$\max_{\beta \geq 0} F_n(\beta; w) = \hat{\mathbb{E}}^w_{B-W}\left[f\left(\|X - X'\|^q_{q,\beta}\right)\right] - \lambda \|\beta\|_1$$

We can make it easier for gradient ascent to find non-zero stationary points.

Claim: $q = 1$ is *better* than $q = 2$.

Reason:

- The gradient itself contains more statistical information when $q = 1$!
- (Sometimes) 0 not stationary when $q = 1$ but is stationary when $q = 2$.

Why $q = 1$ is Better than $q = 2$

- The gradient itself contains more statistical information when $q = 1$!
- (Sometimes) $q = 1$ makes 0 not a stationary point!

Example

Assume $S = \{1\}$ so that $X_1 \not\perp Y$. We want $\beta_1 > 0$.

Start from $\beta = 0$. Compute the gradient w.r.t β_1 at $\beta = 0$.

$$\frac{\partial}{\partial \beta_1} F(\beta) \mid_{\beta=0} = f'(0) \cdot \mathbb{E}_{B-W}\left[|X_1 - X_1'|^q\right]$$

- Key: $\beta = 0$ can never be a stationary point when $q = 1$!

$$\text{Reason}: \quad \frac{\partial}{\partial \beta_1} F(\beta) \mid_{\beta=0} \; \propto \; \mathbb{E}_{B-W}\left[|X_1 - X_1'|\right] > 0.$$

- Key: $\beta = 0$ can be a (bad) stationary point when $q = 2$!

$$\text{Reason}: \frac{\partial}{\partial \beta_1} F(\beta) \mid_{\beta=0} \; \propto \; \mathbb{E}_{B-W}\left[|X_1 - X_1'|^2\right] \stackrel{?}{=} 0.$$

Recovery of Main Effects

Theorem: $n \sim \log p$ samples for recovery of main effects

Let $S = \{1, \ldots s\}$ and $X_1 \perp X_2 \perp \ldots \perp X_s | Y$. Assume

$$\min_{1 \leq j \leq S} \mathbb{E}_{B-W}\left[|X_j - X_j'|\right] \gtrsim \lambda = \Omega\left(\sqrt{\frac{\log p}{n}}\right).$$

Then $\hat{S} = S$ w.h.p. *Note*: $\mathbb{E}_{B-W}\left[|X_j - X_j'|\right] = 0$ if and only if $X_j \perp Y$.

Proof Sketch:

- $\beta = 0$ is not a stationary point.
- Conditional independence implies that reweighting does not affect signal of unselected variables. Rinse and Repeat.

Recovery of Pure Interaction

Theorem: $n \sim p^{2(s-1)} \log p$ samples for recovery of pure interaction

Let X_S be a pure interaction. Let gradient ascent be initialized at $\beta_j \asymp \frac{1}{p}$. Assume

$$\mathbb{E}_{B-W}\left[f\left(\|X_S - X'_S\|_1\right)\right] \gtrsim \sqrt{\frac{p^{2(s-1)} \log p}{n}}.$$

Then $\hat{S} = S$ w.h.p. *Note*: $\mathbb{E}_{B-W}\left[f(\|X_S - X'_S\|_1)\right] = 0$ if and only if $X_S \perp Y$.

Proof Sketch:

- $\beta = 0$ is a bad stationary point in pure interaction case (for both $q = 1, 2$).
- The key is to show the gradient ascent iterates are bounded away from 0 (in the case $q = 1$):

$$\beta_S^{(k)} \gtrsim \frac{1}{p}\mathbf{1}_S \text{ for all iteration } k \in \mathbb{N}.$$

Recovery of Pure Interaction

Theorem: $n \sim p^{2(s-1)} \log p$ samples for recovery of pure interaction

Let X_S be a pure interaction. Let gradient ascent be initialized at $\beta_j \asymp \frac{1}{p}$. Assume

$$\mathbb{E}_{B-W}\left[f\left(\|X_S - X'_S\|_1\right)\right] \gtrsim \sqrt{\frac{p^{2(s-1)} \log p}{n}}.$$

Then $\hat{S} = S$ w.h.p. *Note*: $\mathbb{E}_{B-W}\left[f(\|X_S - X'_S\|_1)\right] = 0$ if and only if $X_S \perp Y$.

Statistics vs. Computation Tradeoff

- Computation cost: $O(p)$ $\Leftrightarrow$ Sample complexity: $n \sim O(p^{2(s-1)\log p})$.
- Computation cost: $O(p^k)$ $\Leftrightarrow$ Sample complexity: $n \sim O(p^{2(s-k)_+ \log p})$.

Recovery of Hierarchical Interaction

Example

(X_1, X_2) is a hierarchical interaction if

- X_1 is dependent of Y, while X_2 is not.
- X_2 is dependent of Y, when conditional on X_1.

Higher order generalizations are possible.

Definition (Hierarchical Interaction (Liu and R. 20'))

The variables in S interacts hierarchically if there exists a nested sequence

$$\emptyset = S_0 \subsetneq S_1 \subsetneq S_2 \ldots \subsetneq S_s = S$$

such that

- $X_{S_k \setminus S_{k-1}}$ is dependent of Y given $X_{S_{k-1}}$.
- $X_{S \setminus S_k} \perp Y \mid X_A$ for any subset $A \subsetneq S_k$.

Recovery of Hierarchical Interaction

Theorem: $n \sim \log p$ samples for recovery of hierarchical interaction

Let X_S be a hierarchical interaction with nested sequence

$$\emptyset = S_0 \subsetneq S_1 \subsetneq S_2 \ldots \subsetneq S_s = S.$$

Then, $\hat{S} = S$ w.h.p. if the following condition holds:

$$\min_{1 \le k \le s} \mathbb{E}_{B-W}^{(w_{S_{k-1}})}\left[f\left(\left\|X_{S_k} - X'_{S_k}\right\|_1\right)\right] \gtrsim \lambda = \sqrt{\frac{\log p}{n}}$$

Note: $\mathbb{E}_{B-W}^{(w_{S_{k-1}})}\left[f(\left\|X_{S_k} - X'_{S_k}\right\|_1)\right] = 0$ if and only if $X_{S_k \setminus S_{k-1}} \perp Y \mid X_{S_{k-1}}$

Proof Sketch:

- $\beta = 0$ is not a stationary point.

Toolkit: Fourier Analysis

Understand the gradient of $F(\beta)$ near 0

$$F(\beta) = \mathbb{E}_{B-W}\left[f(\|X - X'\|_{q,\beta}^{q})\right]$$

- For simplicity, assume $f(x) = \exp(-x)$ in below discussion.

A representation: (idea traced back to Bochner, Herglotz...)

$$F(\beta) = \int |\phi_0(\omega) - \phi_1(\omega)|^2 \prod_j q_\beta(\omega_j) d\omega.$$

where $\phi_y(\omega) = \mathbb{E}[e^{i\langle \omega, X\rangle} \mid Y = y]$ is the characteristic function of $\mathcal{L}(X \mid Y = y)$.

- The function $q_\beta(\omega) = \frac{1}{\pi}\frac{\beta\omega}{\omega^2+\beta^2}$ is the *Cauchy* density of scale β when $q = 1$. Note: *Cauchy* is the Fourier transform of *Laplace* $f(|x|) = \exp(-|x|)$.
- The function $q_\beta(\omega) = \frac{1}{\sqrt{2\pi}\beta}\exp(-\frac{\omega^2}{\beta^2})$ is the *Gaussian* density when $q = 2$. Note: *Gaussian* is the Fourier transform of *Gaussian* $f(|x|^2) = \exp(-|x|^2)$.

Toolkit: Fourier Analysis

Understand the gradient of $F(\beta)$ near $0 \Leftrightarrow$ how fast $F(\beta)$ grows away from 0

$$F(\beta) = \int |\phi_0(\omega) - \phi_1(\omega)|^2 \prod_j q_\beta(\omega_j) d\omega.$$

Recall $q_\beta(\omega) = \frac{\beta\omega}{\omega^2+\beta^2}$ is the Cauchy density ($q = 1$).

Key property: Cauchy density $q_\beta(\omega)$ is *self-bounding* w.r.t β:

$$\frac{q_\beta(\omega)}{q_{\beta'}(\omega)} \geq \frac{\beta}{\beta'} \text{ whenever } \beta \leq \beta'.$$

An Application: Hence $F(\beta)$ is also *self-bounding* when $q = 1$. In particular,

$$F(\beta) \gtrsim F(\mathbf{1}) \cdot \prod_j \beta_j.$$

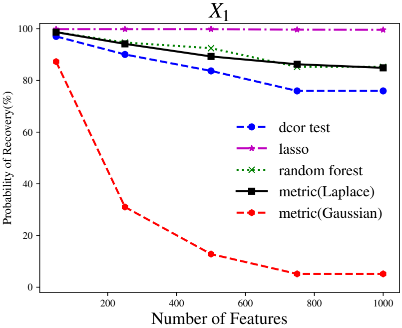
Note $F(\mathbf{0}) = 0$. This gives a crude bound on the gradient that holds for all type of signals:

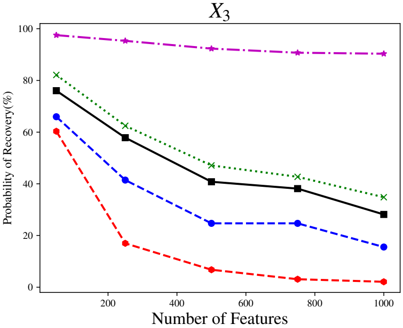
$$\partial_{\beta_k} F(\beta) \gtrsim F(\mathbf{1}) \cdot \prod_{j \neq k} \beta_j.$$

Example: Recovery of Main Effects

$$X_j \mid Y = 0 \sim \mathsf{N}(0, \sigma^2(1+\delta_j)) \text{ for } j = 1, 3.$$

$$\delta_1 = 0.4 \text{ and } \delta_3 = 0.3$$



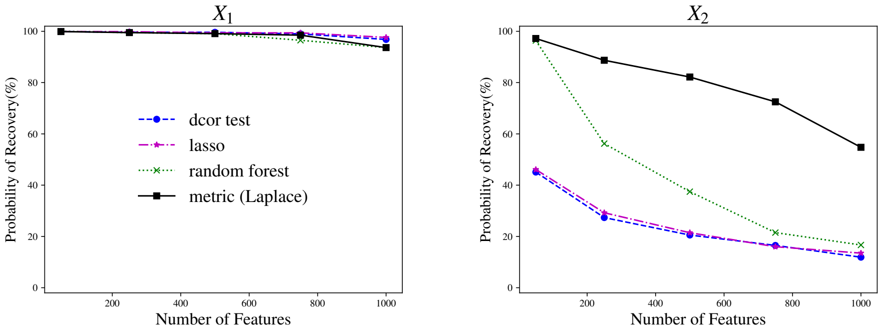


Conclusion:

- $q = 1$ (Laplace) does strictly better than $q = 2$ (Gaussian).
- RF (Random Forest) is the winner, slightly better than MS (Metric screening).

Recovery of Hierarchical Effects (QDA Model)

$$(X_1, X_2) \mid Y = \pm 1 \sim \mathsf{N}\left(\begin{pmatrix} \pm 0.25 \\ \pm 0.1 \end{pmatrix}, \begin{pmatrix} 1, & \pm 0.5 \\ \pm 0.5, & 1 \end{pmatrix}\right).$$



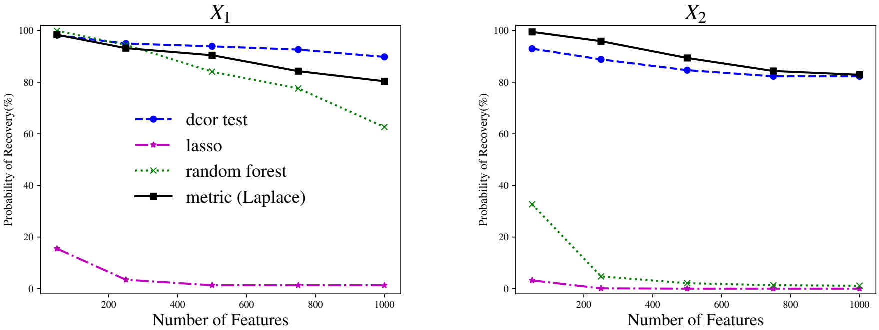
Conclusion:

- Dcor test and Lasso perform poorly in detecting weak main effect signal.
- MS (Metric screening) is the winner, and it scales better in high dimension.

Recovery of Ratio Interaction Signal

$$\text{logit}\, \mathbb{P}(Y = 1|X) = \frac{|X_2|}{|X_1|}.$$

X_1 is the stronger main effect and X_2 is the weaker main effect.

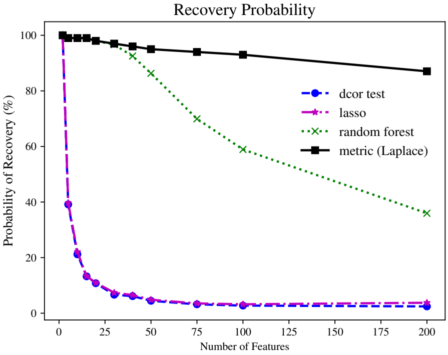


Conclusion:

- MS (Metric screening) is more effective in exploiting interactions than RF.

Recovery of Pure Interaction

$$Y = \begin{cases} +1 & \text{if } X_1 X_2 > 0 \\ -1 & \text{if } X_1 X_2 < 0 \end{cases}$$



Conclusion:

- MS (Metric screening) is clearly the winner of all.

Conclusion (Takeaways)

Detecting interactions is an interesting *combinatorial* problem.

Three Main Ideas:

- Idea 1: Nonparametric two sample test $\Rightarrow$ Maximize dependence measure.
- Idea 2: Inconsistency of naive maximization (masking) $\Rightarrow$ Reweighting.
- Idea 3: Nonconvexity makes it hard to find the global maximum $\Rightarrow$ Design the objective landscape (gradient).