# A Counterfactual Perspective of **Heritability**, **Explainability**, and **ANOVA**

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

January 7, 2026 @ UCL Gatsby Unit

# This talk

Ongoing project, two working papers:

1. **Heritability: A Counterfactual Perspective.** (Joint work with Haochen Lei, Jieru Shi, Hongyuan Cao. Available on my website.)
2. **Counterfactual Explainability and Analysis of Variance.** (Joint work with Zijun Gao. Available on arXiv:2411.01625.)

Haochen Lei
Florida State

Jieru Shi
UCL

Hongyuan Cao
Florida State

Zijun Gao
USC Marshall

# Motivation

We wanted to answer a simple question:

Is there a good notion of **causal variable importance**?

This question is shaped by frustration from three ends:

1. The **causal inference** literature is obssessed with estimating the effect of a specific intervention (particularly the ATE $= \mathrm{E}[Y(1) - Y(0)]$).
2. The **global sensitivity analysis** literature (e.g. functional ANOVA) assumes independent inputs and deterministic outputs.
3. The **genetics** literature is dominated by using linear models to estimate heritability.

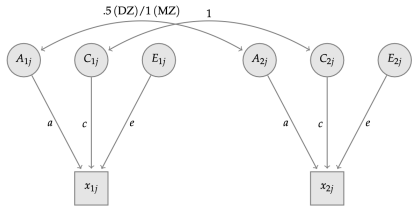
# An exception

- Twin study is a classical method for estimating heritability. **Falconer's formula**:

$$h^2_{\text{twin}} = 2(\rho_{\text{MZ}} - \rho_{\text{DZ}}).$$

  $\rho_{\text{MZ}}/\rho_{\text{DZ}}$ is the correlation of a trait (e.g. height) between monozygotic/dizygotic twins. (The study of correlation between relatives goes back to Fisher's classic 1918 paper.)

- This is often justified using the **ACE model**: $Y = A + C + E$.
- Kohler et al. (2011)[1] formalized this as a linear **structural equation model**:




[1] Social science methods for twins data: Integrating causality, endowments, and heritability. *Biodemography and Social Biology* 2011.

# Our first idea

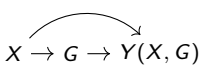
- We made real progress after thinking about the **"fraternal twin counterfactual"**.
- Let $Y(G)$ be the potential outcome of someone with genotype $G$, and $Y(G')$ be the counterfactual of a hypothetical twin with genotype $G'$ that is an i.i.d. copy of $G$.
- We can define **counterfactual heritability** as

$$\xi = \frac{\mathrm{Var}(Y(G) - Y(G'))}{2\,\mathrm{Var}(Y(G))} = \frac{\mathrm{Var}(Y(G,E) - Y(G',E))}{2\,\mathrm{Var}(Y(G,E))}. \quad \text{(Environment } E \textbf{ is shared.)}$$
$$= 1 - \mathrm{Cor}(Y(G), Y(G')).$$

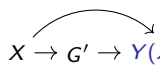
- Not hard to see that under the ACE model, $\xi = h^2_{\text{twin}}/2$. The factor $1/2$ is because $\xi$ only measure heritability in **one generation**.
- In many way, this is a nice definition. A practical challenge is that, without further strong assumptions (e.g. additivity in the ACE model), the value of $\xi$ **cannot be uniquely determined** even with infinite amount of data.

# Distinguishing notions of heritability by the counterfactual comparison

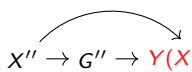
- Geneticists talk about **broad-sense heritability**, **narrow-sense heritability**, and **SNP heritability**. Often only defined under specific models and are tricky to interpret.
- Let $X$ be parent genotypes and $G$ be children genotypes.



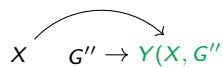

(a) Potential outcomes.




(b) Fraternal counterfactual.




(c) Unrelated counterfactual.




(d) Adopted counterfactual.

(b) **Fraternal**: "one-generation" heritability $\propto \mathrm{Var}\{Y(X, G) - Y(X, G')\}$.

(c) **Unrelated**: "infinite-generation" heritability $\propto \mathrm{Var}\{Y(X, G) - Y(X', G'')\}$.

(d) **Adopted**: heritability $\propto \mathrm{Var}\{Y(X, G) - Y(X, G'')\}$.

- Implicit in relatedness disequilibrium regression (Young et al., 2018)[2] but not always $\leq 1$.

[2] Relatedness disequilibrium regression estimates heritability without environmental bias. *Nature Genetics.*

# Partial identification (Theorem 2 in Lei, Shi, Cao, Zhao)

Suppose $G \perp\!\!\!\perp \{Y(g) : g \in \mathcal{G}\} \mid X$. Then

$$\xi_l' \le \xi_l \le \xi \le \min\{\xi_u, \xi_u'\},$$

where

$$\xi_l' = \frac{\mathrm{E}[\mathrm{Var}\{\mathrm{E}(Y \mid G, X) \mid X\}]}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{E}(\mathrm{Var}(Y \mid G, X)) + \mathrm{Var}(\mathrm{E}(Y \mid X))}{\mathrm{Var}(Y)},$$

$$\xi_l = \frac{\mathrm{E}\left[\left\{F_{G,X}^{-1}(U) - F_{G',X}^{-1}(U)\right\}^2\right]}{2\,\mathrm{Var}(Y)}, \qquad \xi_u = \frac{\mathrm{E}\left[\left\{F_{G,X}^{-1}(U) - F_{G',X}^{-1}(1-U)\right\}^2\right]}{2\,\mathrm{Var}(Y)},$$

$$\xi_u' = \frac{\mathrm{E}(\mathrm{Var}(Y \mid X))}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Var}(\mathrm{E}(Y \mid X))}{\mathrm{Var}(Y)}.$$

- $F_{G,X}^{-1}(\cdot)$ is the conditional quantile function of $Y$ given $G, X$. $U \sim \mathrm{Unif}[0,1]$.
- The lower bound $\xi_l$ is tight. The upper bound $\xi_u$ is tight when $G$ is binary but generally not very useful.

# Comparison with other notions of heritability

- Let $H^2 = \dfrac{\mathrm{Var}(\mathrm{E}(Y \mid G))}{\mathrm{Var}(Y)}$ and $h^2 = \dfrac{\mathrm{Var}(G^T\theta)}{\mathrm{Var}(Y)}$ ($\theta$ is the least squares projection).
- $G_1, G_2 \sim \mathrm{Bern}(0.5)$, $X, E_1, E_2 \sim \mathrm{N}(0, 0.25)$, **all independent**.

| $Y(g) =$ | Narrow $h^2$ | Broad $H^2$ | Counterfactual $\xi$ | Lower $\xi_l'$ | Tight lower $\xi_l$ | Upper $\xi_u'$ |
|---|---|---|---|---|---|---|
| $\beta_1 g_1 + \beta_2 g_2 + E_1$ | $\frac{\beta_1^2+\beta_2^2}{1+\beta_1^2+\beta_2^2}$ | $\frac{\beta_1^2+\beta_2^2}{1+\beta_1^2+\beta_2^2}$ | $\frac{\beta_1^2+\beta_2^2}{1+\beta_1^2+\beta_2^2}$ | $\frac{\beta_1^2+\beta_2^2}{1+\beta_1^2+\beta_2^2}$ | $\frac{\beta_1^2+\beta_2^2}{1+\beta_1^2+\beta_2^2}$ | 1 |
| $\beta g_1 g_2 + E_1$ | $\frac{2\beta^2}{4+3\beta^2}$ | $\frac{3\beta^2}{4+3\beta^2}$ | $\frac{3\beta^2}{4+3\beta^2}$ | $\frac{3\beta^2}{4+3\beta^2}$ | $\frac{3\beta^2}{4+3\beta^2}$ | 1 |
| $\beta g_1 E_2 + E_1$ | 0 | 0 | $\frac{\beta^2}{4+2\beta^2}$ | 0 | $\frac{(\sqrt{1+\beta^2}-1)^2}{4+2\beta^2}$ | 1 |
| $\beta_1 g_1 + \beta_2 X + E_1$ | $\frac{\beta_1^2}{1+\beta_1^2+\beta_2^2}$ | $\frac{\beta_1^2}{1+\beta_1^2+\beta_2^2}$ | $\frac{\beta_1^2}{1+\beta_1^2+\beta_2^2}$ | $\frac{\beta_1^2}{1+\beta_1^2+\beta_2^2}$ | $\frac{\beta_1^2}{1+\beta_1^2+\beta_2^2}$ | $\frac{1+\beta_1^2}{1+\beta_1^2+\beta_2^2}$ |
| $\beta g_1 X + E_1$ | 0 | 0 | $\frac{\beta^2}{4+2\beta^2}$ | $\frac{\beta^2}{4+2\beta^2}$ | $\frac{\beta^2}{4+2\beta^2}$ | $\frac{4+\beta^2}{4+2\beta^2}$ |

# Our second idea

- Let $\xi(G)$ be the counterfactual heritability above. Can similarly define $\xi(E)$ and $\boldsymbol{\xi(G \vee E)}$.
- Row 3 above shows that $G$ receives "credit" from the interaction term $G \times E$.
- Can we define **interaction explainability** using the inclusion-exlusion principle as

$$\xi(G \wedge E) = \xi(G) + \xi(E) - \xi(G \vee E)?$$

- It turns out that this can be rewritten as

$$\xi(G \wedge E) = \frac{\mathrm{Var}(Y(G,E) - Y(G',E) - Y(G,E') + Y(G',E'))}{4\,\mathrm{Var}(Y)} \geq 0.$$

  The **second-order finite difference** has indeed been used to describe interaction.
- We wrote a conference paper based on this and shared it with Professor Art Owen. He pointed out that this is the **superset importance** introduced by Giles Hooker[3] with a variance-based formula in Liu and Owen (2006)[4].

[3] Discovering additive structure in black box functions. KDD 2004.

[4] Estimating mean dimensionality of analysis of variance decompositions. *JASA* 2006.

# Related literature: Functional ANOVA

- Let $f$ be a function of independent random variables $W_1, \ldots, W_K$.
- We can always write $f(W)$ as

$$f(W) = \sum_{\mathcal{S} \subseteq [K]} f_{\mathcal{S}}(W_{\mathcal{S}}),$$

where the terms can be obtained inductively by (let $f_{\emptyset}(W) = \mathsf{E}[f(W)]$)

$$f_{\mathcal{S}}(w_{\mathcal{S}}) = \mathsf{E}\left[f(W) - \sum_{\mathcal{S}' \subset \mathcal{S}} f_{\mathcal{S}'}(W) \mid W_{\mathcal{S}} = w_{\mathcal{S}}\right], \quad \text{for } \mathcal{S} \subseteq [K].$$

- The terms are orthogonal: $\mathsf{E}[f_{\mathcal{S}}(W_{\mathcal{S}}) f_{\mathcal{S}'}(W_{\mathcal{S}'})] = 0$ for all different $\mathcal{S}, \mathcal{S}' \subseteq [K]$.
- So we have the **functional ANOVA** decomposition:

$$\mathrm{var}(f(W)) = \sum_{\mathcal{S} \subseteq [K]} \sigma_{\mathcal{S}}^2, \quad \text{where } \sigma_{\mathcal{S}}^2 := \mathrm{var}(f_{\mathcal{S}}(W_{\mathcal{S}})).$$

- This goes back to Hoeffding (1948, *AOMS*).

# Related literature: global sensitivity analysis

Some notions of variable importance for a subset $\mathcal{S} \subseteq [K]$:

- **Sobol's lower and upper sensitivity indices**: $\underline{\tau}^2_{\mathcal{S}} = \sum_{\mathcal{S}' \subseteq \mathcal{S}} \sigma^2_{\mathcal{S}'}$ and $\overline{\tau}^2_{\mathcal{S}} = \sum_{\mathcal{S}' \cap \mathcal{S} \neq \emptyset} \sigma^2_{\mathcal{S}'}$.
- **Super-set importance**: $\overline{\sigma}^2_{\mathcal{S}} = \sum_{\mathcal{S}' \supseteq \mathcal{S}} \sigma^2_{\mathcal{S}'}$.
- **Shapley value** (not going into details...).

## Pick-freeze method

The following formulas are used to accelerate GSA:

$$\underline{\tau}^2_{\mathcal{S}} = \mathrm{Cov}(f(W), f(W_{\mathcal{S}}, W'_{-\mathcal{S}})) \quad \text{and} \quad \overline{\tau}^2_{\mathcal{S}} = \frac{1}{2}\,\mathrm{E}[\{f(W) - f(W'_{\mathcal{S}}, W_{-\mathcal{S}})\}^2],$$

$$\overline{\sigma}^2_{\mathcal{S}} = 2^{-|\mathcal{S}|}\mathrm{var}\{I_{\mathcal{S}}(W, W')\}, \quad \text{for all } \mathcal{S} \subseteq [K],$$

where $I_{\mathcal{S}}(w, w') = \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} f(w'_{\mathcal{S}'}, w_{-\mathcal{S}'})$ is an interaction contrast (which forms the **anchored decomposition**, not going into details...)

# Unified perspective: explainability as a probability measure

- The **explanation algebra** $\mathcal{E}(W)$ is the Boolean algebra generated by $W_1, \ldots, W_K$ using conjunction $\vee$, disjunction $\wedge$, and negation $\neg$.

## Theorem 1 in Gao and Zhao

Let $\xi_1, \xi_2, \xi_3, \xi_4$ be any probability measures on $\mathcal{E}(W)$ such that, for all $\mathcal{S} \subseteq [K]$,

$$\xi_1\left((\wedge_{k\in\mathcal{S}} W_k) \wedge (\wedge_{k\notin\mathcal{S}} \neg W_k)\right) = \frac{\sigma^2_{\mathcal{S}}}{\mathrm{var}(f(W))},$$

$$\xi_2\left(\neg(\vee_{k\notin\mathcal{S}} W_k)\right) = \frac{\underline{\tau}^2_{\mathcal{S}}}{\mathrm{var}(f(W))},$$

$$\xi_3(\vee_{k\in\mathcal{S}} W_k) = \frac{\overline{\tau}^2_{\mathcal{S}}}{\mathrm{var}(f(W))},$$

$$\xi_4(\wedge_{k\in\mathcal{S}} W_k) = \frac{\overline{\sigma}^2_{\mathcal{S}}}{\mathrm{var}(f(W))}.$$

Then $\xi_1 = \xi_2 = \xi_3 = \xi_4$.

# Implications

This is essentially a re-formulation of functiona ANOVA, but we can use the familiar probability theory to derive many implications:

1. $\xi(\vee_{k\in\mathcal{S}} W_k) = 1$.
2. $\xi(\vee_{k\in\mathcal{S}'} W_k) \geq \xi(\vee_{k\in\mathcal{S}} W_k)$ and $\xi(\wedge_{k\in\mathcal{S}'} W_k) \leq \xi(\wedge_{k\in\mathcal{S}} W_k)$ for any $\mathcal{S} \subseteq \mathcal{S}' \subseteq [K]$.
3. $\xi(W_1) + \ldots \xi(W_K) \geq \xi(\vee_{k\in[K]} W_k)$.

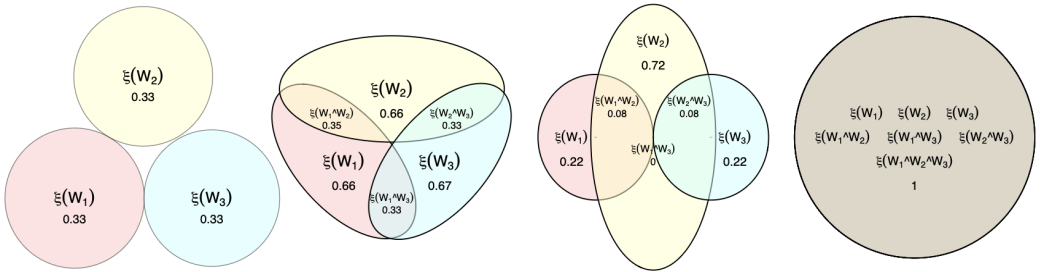
## Remarks

- 2.2 is what motivated Hooker's **superset-importance** (name is unhelpful...).
- 3 is the **Efron-Stein inequality**

$$\text{Var}(f(W)) \leq \sum_{k=1}^{K} \text{Var}(f(W) - f(W'_k, W_{-k})).$$

  This can be improved by the Boole-Bonferroni inequalities.
- For practice, the important thing is that we can use **Venn's diagram** to visualize $\xi$.

# Examples



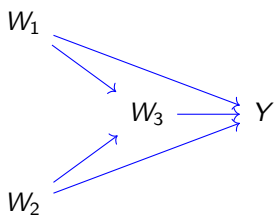

Left to right:

1. Linear: $f(w) = w_1 + w_2 + w_3$;
2. Quadratic: $f(w) = w_1 w_2 + w_1 w_3 + w_2 w_3$;
3. Single-layer NN (sigmoid activation): $f(w) = (1 + e^{10w_1 + 10w_2})^{-1} + (1 + e^{10w_2 + 10w_3})^{-1}$;
4. Multilinear monomial: $f(w) = w_1 w_2 w_3$.

# Third idea

- How to extend this to dependent explanations?
- Our idea is to use a directed acyclic graph (DAG) to model their causal dependence and **define "fraternal counterfactuals" by resampling the intrinsic noise**.

## Example

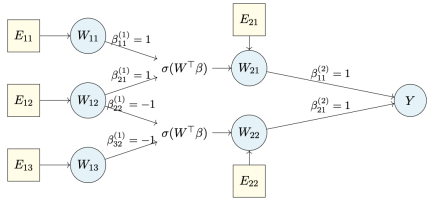



Pearl's NPSEM-IE assumes (in potential outcomes terms)

$$
\begin{aligned}
W_1 &= f_1(E_1), \\
W_2 &= f_2(E_2), \\
W_3(w_1, w_2) &= f_3(w_1, w_2, E_3), \\
Y(w_1, w_2, w_3) &= f_4(w_1, w_2, w_3, E_4),
\end{aligned}
$$

- By recursive substitution, we can write $Y$ as a function of the intrinsic noises $E_1, E_2, E_3, E_4$, which are assumed to be independent in Pearl's model.
- Janzing et al. (2024, AISTATS) suggested the same idea for causal extensions of general dependence measures (e.g. mutual information) in a more conceptual paper.

# Toy example

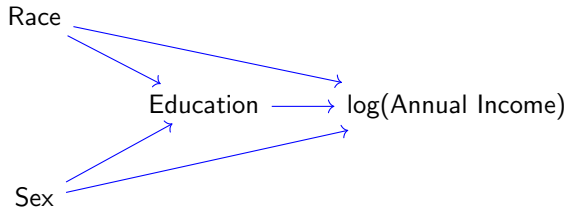



(a) Network structure.


(b) Equal noise variance.

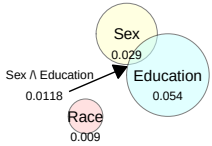
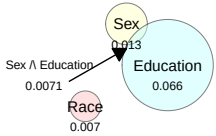
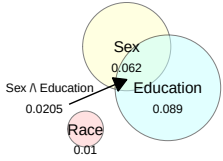
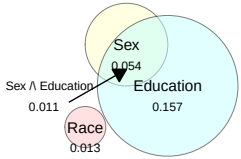
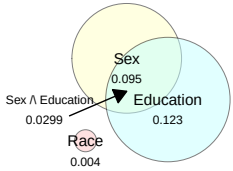
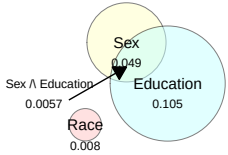

(c) Second layer noise variance decreased.

# Real data example: Explaining income inequality

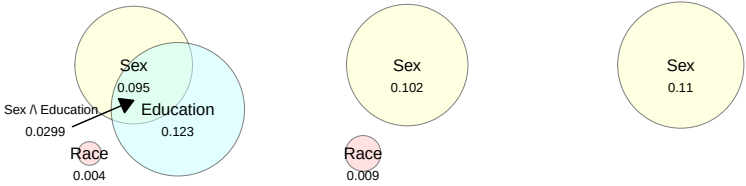



- Datasets: UCI `Adult` (1994) and `ACSIncome` (2018).
- We assumed the basic potential outcomes are **comonotone ($\Rightarrow$ point identification)** and sampled the counterfactuals after estimating the conditional distributions using XGBoost.

# Results

| Age | Income in 1994 | Income in 2018 |
|---|---|---|
| $[25, 30)$ |  |  |
| $[40, 45)$ |  |  |
| Over 60 |  |  |

# Consistency across graphs




- Counterfactual explainability generally depends on the causal DAG but has certain consistency properties with respect to finer mechanistic explanations (see the paper for some formal results).
- In this example, Sex should have the same total explainability regardless of whether Race and Education are included.

# Conclusion

The papers have more theorems (inc. an axiomatization of $\xi$), examples, and discussion.

## Why is counterfactual heritability/explainability a good idea?

### In theory

It provides a counterfactual extension to functional ANOVA that can be applied with dependent explanations and non-deterministic outputs. Has certain consistency properties.

### In practice

Uses contextual mechanistic information. Easy to visualize and interpret.

**Thank you!**