

# Purifying Interaction Effects with Functional ANOVA: Paper Walkthrough + Two Experiments

Aik Tarkhanyan (LMU, Munich)

[https://github.com/HaykTarkhanyan/fanova\\_purification](https://github.com/HaykTarkhanyan/fanova_purification)

# Agenda

## Part I: Paper Walkthrough

1. The identifiability problem
2. Functional ANOVA target
3. Mass-moving algorithm
4. Convergence & properties
5. Weight estimation ( $w$ )

## Part II: Experiments

6. German Credit (real data)  
Main effects & interactions under different  $w$
7. Stress-test (synthetic)  
How dependence  $\rho$  affects purification

## Conclusions

Summary & takeaways

# Agenda

## Part I: Paper Walkthrough

- ▶ 1. The identifiability problem
- 2. Functional ANOVA target
- 3. Mass-moving algorithm
- 4. Convergence & properties
- 5. Weight estimation ( $w$ )

## Part II: Experiments

- 6. German Credit (real data)  
Main effects & interactions under different  $w$
- 7. Stress-test (synthetic)  
How dependence  $\rho$  affects purification

## Conclusions

Summary & takeaways

# 1) Additive models with interactions

We consider models of the form:

$$Y \approx f_0 + f_1(X_1) + f_2(X_2) + f_3(X_1, X_2)$$

More generally:

$$F(X) = f_0 + \sum_{i=1}^d f_i(X_i) + \sum_{i \neq j} f_{ij}(X_i, X_j) + \dots$$

- Goal: interpret  $f_i$  (mains) and  $f_{ij}$  (interactions).
- Problem: without constraints, the decomposition is **non-identifiable**.

## 2) Identifiability issue: effects can move across orders

For a fixed predictor, you can shift “mass” between terms while keeping predictions identical.

### Simple intuition

Add any  $h(X_1)$  to  $f_1$  and subtract the same  $h(X_1)$  from  $f_{12}(X_1, X_2)$ : the sum stays unchanged.

- This allows **contradictory interpretations** for the exact same function.
- So “main effect” and “interaction” are not uniquely defined by the predictor.
- This motivates a **canonical** notion of “pure” interactions.

### 3) Boolean example (OR/AND as XOR + mains)

Let  $X_1, X_2 \in \{0, 1\}$ .

#### Key equivalence

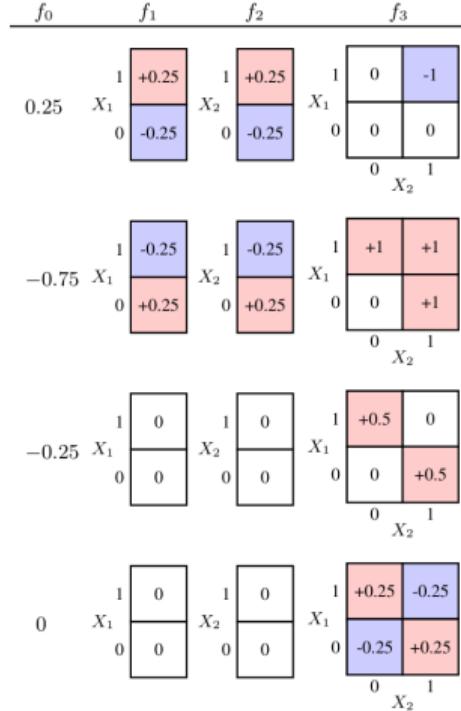
$$X_1 \vee X_2 = 0.25(X_1 \oplus X_2) + 0.5(X_1 - 0.5) + 0.5(X_2 - 0.5) + 0.75$$

Similarly:

$$X_1 \wedge X_2 = -0.25(X_1 \oplus X_2) + 0.5(X_1 - 0.5) + 0.5(X_2 - 0.5) + 0.25$$

- The interaction parts are (centered) XOR up to sign; the rest is absorbed into mains+intercept.
- AND and OR have identical interaction structure (centered XOR up to sign); differences are absorbed into mains/intercept.

# Representational degeneracy



**Takeaway:** Multiple decompositions can yield identical predictions but assign credit differently across mains/interactions; purification selects a canonical one.

## 4) Multiplicative model and the $(\alpha, \beta)$ degree of freedom

A classic interaction model:

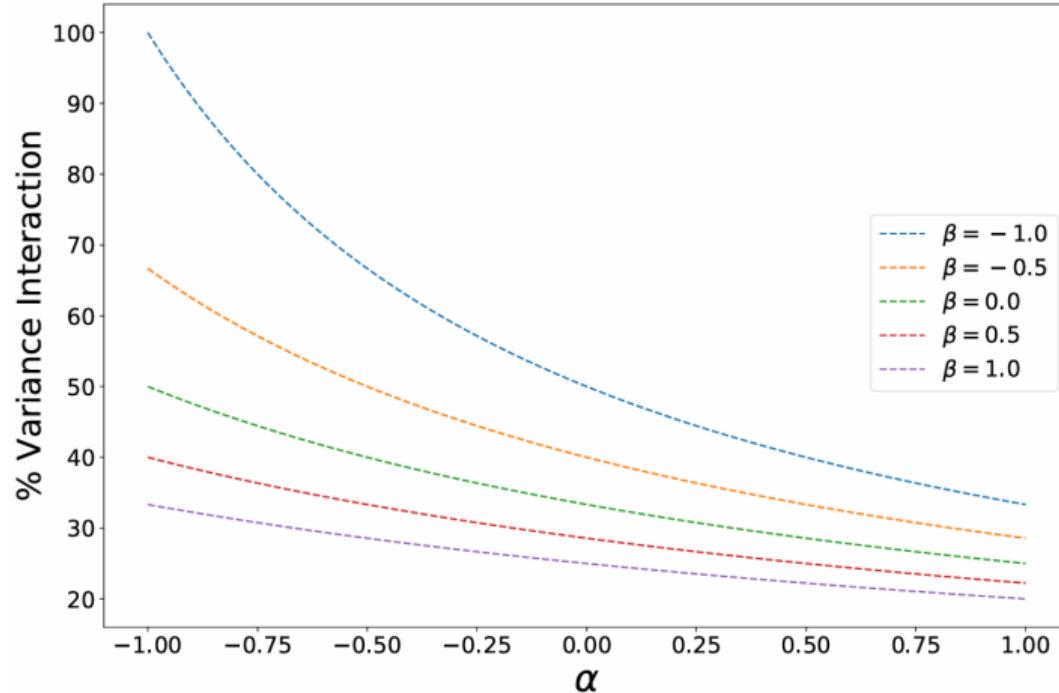
$$Y \approx a + bX_1 + cX_2 + dX_1X_2.$$

An equivalent re-parameterization (for any  $\alpha, \beta$ ):

$$Y \approx (a - d\alpha\beta) + (b + d\beta)X_1 + (c + d\alpha)X_2 + d(X_1 - \alpha)(X_2 - \beta).$$

- Identifiable as a function, but coefficients are **not meaningful** without rules for choosing  $\alpha, \beta$ .
- A canonical choice is the fANOVA one, which yields **minimum variance in higher-order terms**.
- Conceptually, centering constants align with moments like  $\mathbb{E}_w[X_1]$  and  $\mathbb{E}_w[X_2]$ ; the decomposition depends on  $w$ .

## Multiplicative example: varying $(\alpha, \beta)$



**Takeaway:** As  $(\alpha, \beta)$  vary, variance attribution shifts between mains and interactions even though predictions do not.

## 5) Functional ANOVA (fANOVA): the canonical target

The functional ANOVA decomposes  $F$  into **orthogonal** components under weight  $w(X)$ :

$$F(X) = \underbrace{f_0}_{\text{predictor}} + \underbrace{\sum_i f_i(X_i)}_{\text{main effects}} + \underbrace{\sum_{i < j} f_{ij}(X_i, X_j)}_{\text{pairwise interactions}} + \dots$$

**Orthogonality:**  $\mathbb{E}_w[f_u \cdot f_v] = 0$  for  $u \neq v$ .

Equivalently, each component satisfies **integrate-to-zero** (marginalizing out any variable gives 0):

$$\boxed{\mathbb{E}_w[f_u(X_u) | X_{u \setminus i}] = 0 \quad \forall i \in u}$$

- For main effect  $f_i$ :  $\mathbb{E}_w[f_i(X_i)] = 0$  (centered).
- For interaction  $f_{ij}$ :  $\mathbb{E}_w[f_{ij} | X_i] = 0$  and  $\mathbb{E}_w[f_{ij} | X_j] = 0$  (no marginal structure).

## 6) Piecewise-constant case: tensors + slice mean-zero

For piecewise-constant  $F$  on bins (one set per feature), each  $f_u$  becomes a **tensor**  $T_u$ .

**Integrate-to-zero** → **weighted slice mean-zero**:

$$\sum_k T_u[\dots, k, \dots] \cdot w_i[k] = 0 \quad \text{for each slice along dimension } i \in u$$

Pairwise intuition ( $f_{ij}$  is a matrix)

- Each **row** has weighted mean zero:  $\sum_k T_{ij}[r, k] \cdot w_j[k] = 0$
- Each **column** has weighted mean zero:  $\sum_k T_{ij}[k, c] \cdot w_i[k] = 0$

**Takeaway:** This constraint enables an exact post-hoc algorithm for tree ensembles.

# Slice Mean-Zero: Visual Example

Before purification			After purification		
$X_2$	4	2	3	$\bar{r}_3 = 3$	$\bar{r}_3 = 0$
	1	2	3	$\bar{r}_2 = 2$	$\bar{r}_2 = 0$
	1	5	3	$\bar{r}_1 = 3$	$\bar{r}_1 = 0$
$X_1$			$X_1$		
Row means $\neq 0$			Row means $= 0$ ✓		

What happens to the removed mass?

Row means are subtracted from the interaction tensor → added to the main effect  $f_2(X_2)$ .  
Column means are subtracted → added to  $f_1(X_1)$ . Predictions unchanged!

## 7) “Pure interaction effects” = fANOVA of $\mathbb{E}[Y | X]$

**Pure interactions** = variance that *cannot* be explained by any subset of variables.

### Formal definition

Find  $\{f_u\}$  that minimize  $\mathbb{E}_w[(F(X) - \sum_u f_u(X_u))^2]$  subject to:

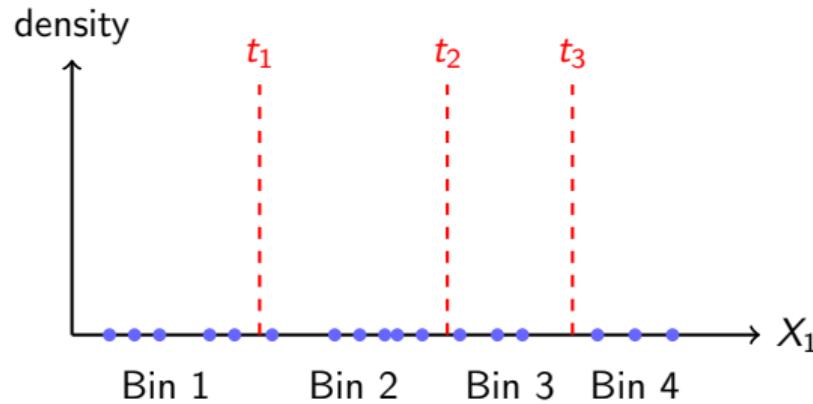
$$\mathbb{E}_w[f_u(X_u) | X_v] = 0 \quad \forall v \subsetneq u$$

### Key insight

This is exactly the fANOVA decomposition with  $w(X) = p(X)$  (true data density).

**Takeaway:** Pure = irreducible. If  $f_{ij}$  can be absorbed into  $f_i$  or  $f_j$ , it's not a “pure” interaction.

## Tree Splits → Bins



**Takeaway:** Tree split thresholds define bins; each bin is a region where  $F$  is constant.

## 8) Purification for trees: mass-moving

For tree ensembles, represent effects as tensors  $T_u$ . **Idea:** iteratively move weighted slice means from higher-order to lower-order tensors.

### One step of mass-moving

- ① Compute weighted mean of slice:  $\bar{m} = \sum_k T_u[\dots, k, \dots] \cdot w_i[k]$
- ② Subtract from interaction:  $T_u[\dots, k, \dots] \leftarrow T_u[\dots, k, \dots] - \bar{m}$
- ③ Add to lower-order tensor:  $T_{u \setminus i}[\dots] \leftarrow T_{u \setminus i}[\dots] + \bar{m}$

### Algorithm:

- **Purify-Matrix:** cycle through dimensions until all slice means  $\approx 0$ .
- **Purify-All:** process tensors from highest-order down to intercept.

# Mass-moving (Purify-Matrix): Visual Walkthrough

## Step 1: Compute row means

2	1	1.5
-1	-2	-1.5
0.5	-0.5	0

$$\begin{aligned}\rightarrow \bar{m}_1 &= 1.5 \\ \rightarrow \bar{m}_2 &= -1.5 \xrightarrow{\text{blue arrow}} \\ \rightarrow \bar{m}_3 &= 0\end{aligned}$$

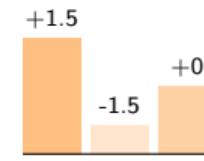
$f_{12}$  (interaction)

## Step 2: Subtract from rows

0.5	-0.5	0
0.5	-0.5	0
0.5	-0.5	0

$f_{12}$  (rows mean 0)

## Step 3: Add to main



$f_2$  (main effect)

### Key insight

Subtracting  $\bar{m}_i$  from row  $i$  of  $f_{12}$  and adding it to  $f_2[i]$  preserves predictions:

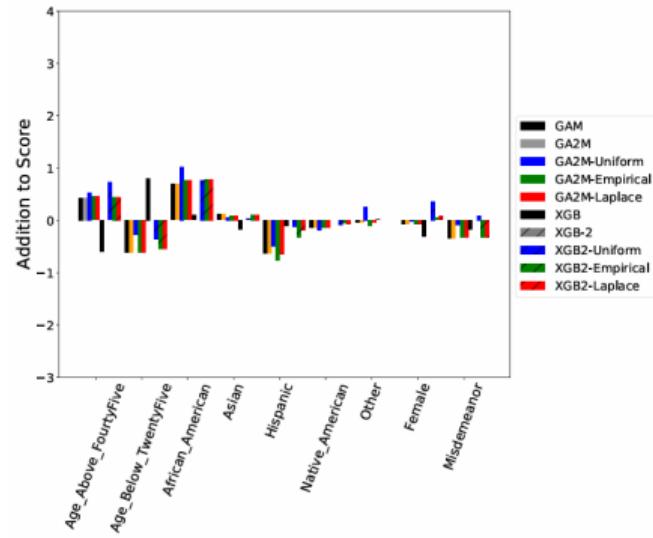
$$f_2(x_2) + f_{12}(x_1, x_2) = \text{unchanged}$$

### Result

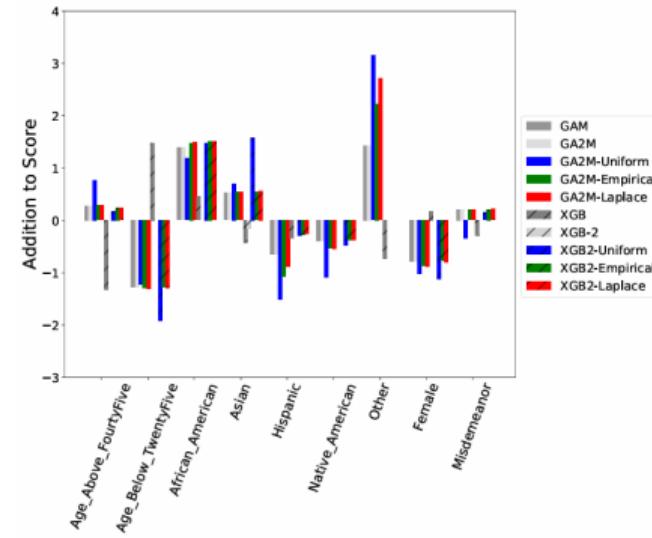
After convergence: every row/column of interaction has weighted mean = 0

**Takeaway:** Mass moves from interaction to mains; predictions stay identical; interaction becomes “pure” (only irreducible structure remains).

# Example: COMPAS + Recidivism Datasets



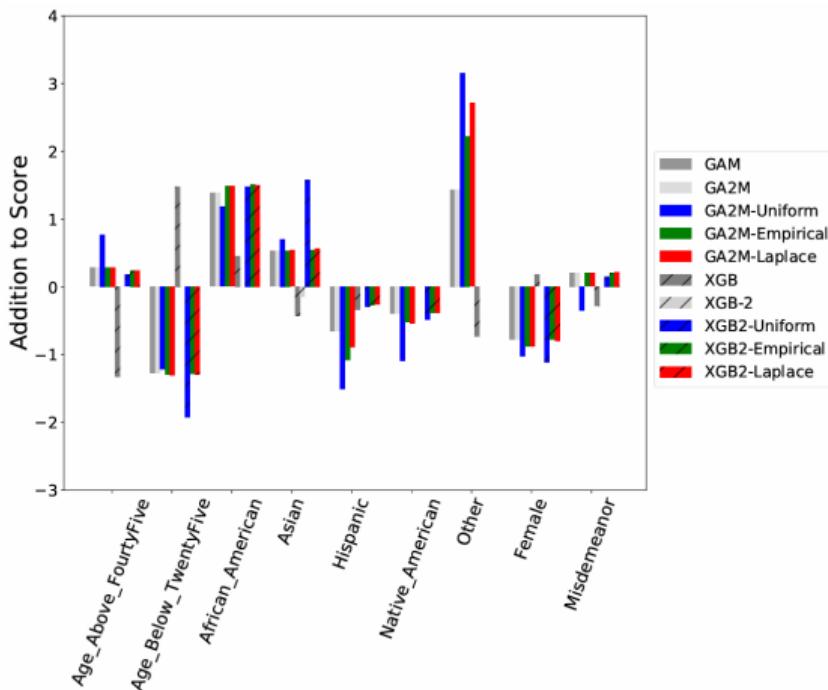
(a) Prediction of Recidivism



(b) Prediction of COMPAS Score

Figure 4: Main effects of additive models with interactions trained to predict the (a) ground-truth recidivism and (b) COMPAS risk score. The implications of the main effects depend on the model class, the use of purification, and the distribution used for purification.

# COMPAS Dataset



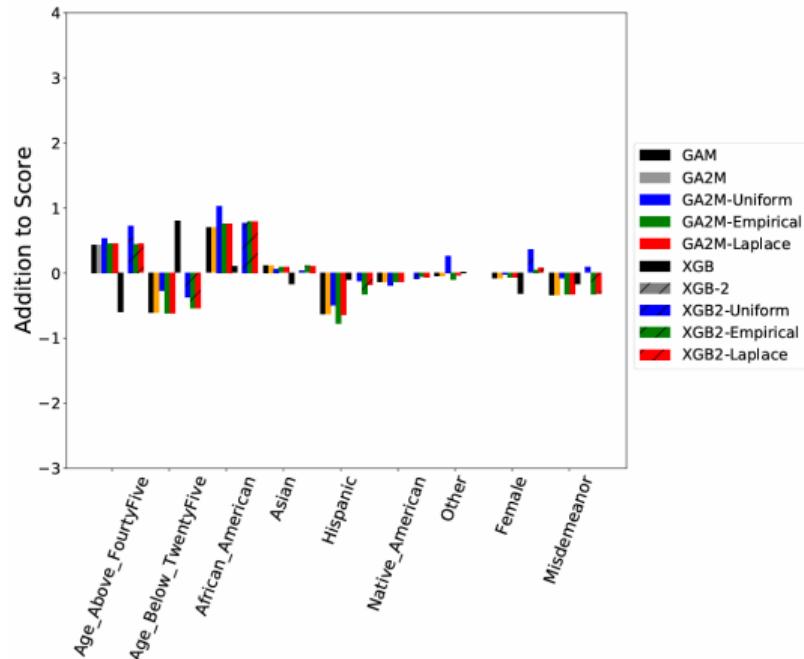
## What is COMPAS?

- Predicts recidivism risk for bail decisions
- High stakes → fair treatment crucial

## Key Insights:

- Main effects **change significantly** after purification
- Magnitude and **sign** depend on weight choice
- XGB signs often **opposite** to purified models

# Recidivism Dataset



## Background:

- Propublica 2016: Broward County, FL
- Includes COMPAS predictions

## Why It Matters:

- Sparked algorithmic bias controversy
- Conflicting conclusions across studies

## Research Question:

Does purification change bias conclusions?

## Example: California housing

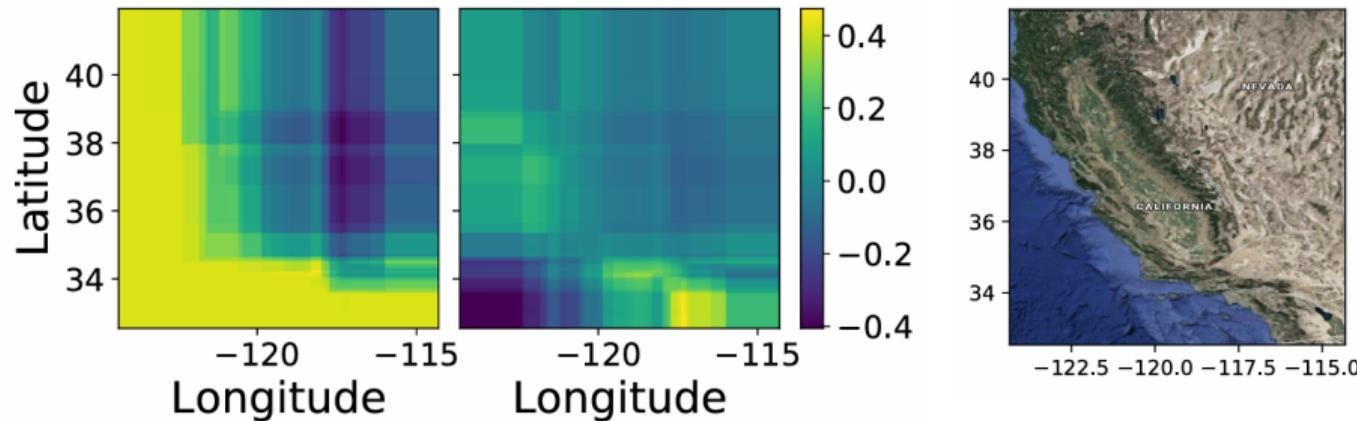


Figure 5: Interaction of the Latitude/Longitude features in an XGB2 model trained on the California housing data. The left pane is the unpurified interaction, the middle pane is the purified interaction, and the right is the map of California from which samples were drawn. Purification sorts out the influence from the Los Angeles and the San Francisco metro areas.

## 10) Convergence + correctness

The mass-moving procedure converges to tensors satisfying slice mean-zero constraints. In practice, a small number of sweeps per interaction tensor is enough (see next slide).

### Theorem 1 (special distributions)

For many simple weights (e.g., uniform along row/column dimensions), purification converges in a single pass.

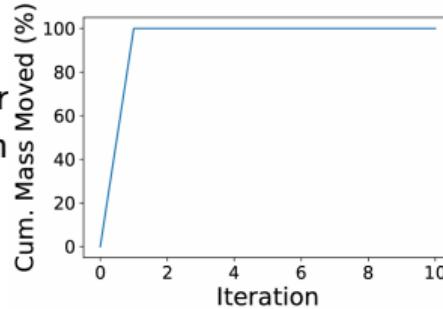
### Theorem 2 (generic non-degenerate $w$ )

Converges to tolerance  $\varepsilon$  in  $O(\log(M_0) - \log(\varepsilon))$  iterations per interaction tensor.

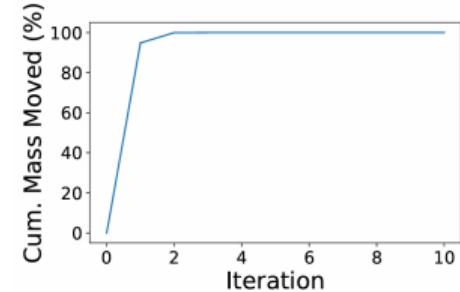
$M_0 = \text{initial unpurified mass } \sum_{i,j} w_{i,j}(r_i + c_j)$ , where  $r_i, c_j$  are row/column means.

# Empirical convergence: mass moves quickly

- Generate random tensors  $T \sim \mathcal{N}(0, \sigma I)$ .
- Use weights either: (i) uniform ( $w \propto 1$ ) or (ii) Gaussian ( $w \sim \mathcal{N}(0, \sigma I)$ ) in dimension  $P$ .
- In practice, almost all mass is moved in the first iteration.
- With uniform weights, convergence occurs in a single pass (per row/column).



(a) Uniform density,  $\sigma = 1$ ,  $P = 100$



(b) Density drawn from multivariate normal  $\sigma = 10$ ,  $P = 100$

**Takeaway:** Most of the mass moves on the first sweep; uniform weights converge in one pass, enabling purification at scale.

## 11) Useful properties

Because the purified decomposition is the unique fANOVA form:

- **Permutation invariance:** reordering categorical codes does not change purified interactions.
- **Linearity:** purification commutes with averaging / bagging (purify ensembles without changing results).

## 12) Estimating $w$ is part of interpretation

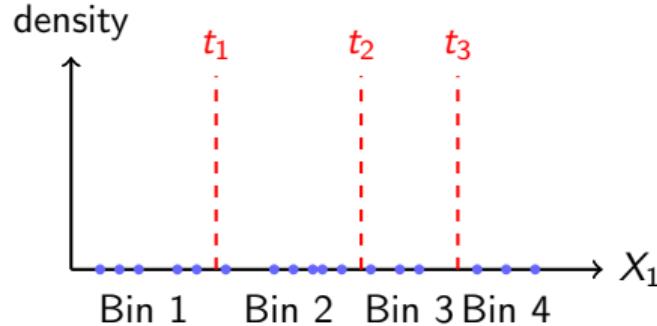
Effects are only meaningful *relative to* a distribution  $w(X)$ . The correct target is the true data density  $p(X)$ , but it must be estimated.

### Three practical estimators (piecewise-constant)

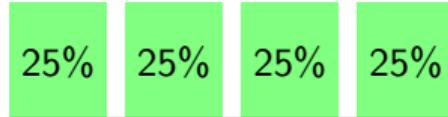
- **Uniform:**  $\hat{w}_{\text{unif}}(x_{-u}) \propto 1$
- **Empirical:**  $\hat{w}_{\text{emp}}(x_{-u}) \propto \sum_{x \in X_{\text{train}}} \mathbf{1}\{x_{-u} = x'_{-u}\}$
- **Laplace:**  $\hat{w}_{\text{lap}} \propto \hat{w}_{\text{unif}} + \hat{w}_{\text{emp}}$  (avoids zero-count bins; stabilizes slice means when support is sparse)

# Binning & Weight Distributions

Tree splits → bins



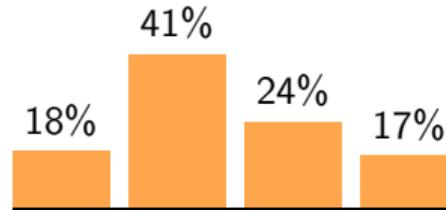
Uniform weights



Weight options

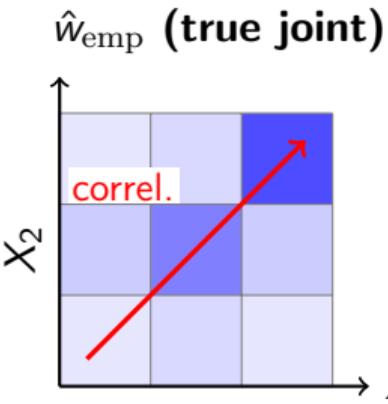
<b>uniform</b>	All bins equal
<b>empirical</b>	Counts from data
<b>laplace</b>	empirical + smooth
<b>indep</b>	$w_{12} = w_1 \cdot w_2$

Empirical weights

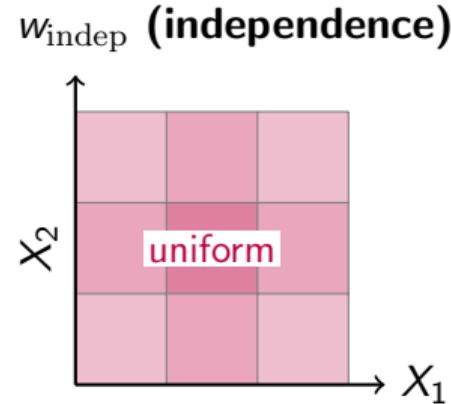


**Takeaway:** Bins derived from tree splits. Weight choice affects the “slice mean zero” constraint.

## 2D Joint Weights: Correlated Features



Mass concentrated on diagonal



$$w_{\text{indep}} = w(x_1) \cdot w(x_2)$$

### Key difference

$\hat{w}_{\text{emp}}$  captures where data lives (diagonal).  $w_{\text{indep}}$  spreads mass uniformly, ignoring correlation.

# Laplace Smoothing: Handling Empty Bins

**Problem:** Some bins may have zero training samples.

- Weighted mean becomes undefined (0/0)
- Rare regions get ignored entirely
- Sensitive to sampling noise

**Solution: Laplace (add-one) smoothing**

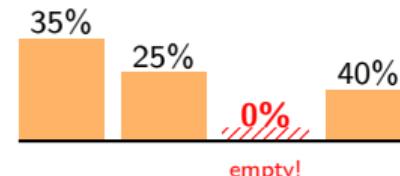
$$\hat{w}_{\text{lap}} = \hat{w}_{\text{emp}} + \hat{w}_{\text{unif}}$$

Equivalently: add a “pseudo-count” of 1 to each bin before computing weights.

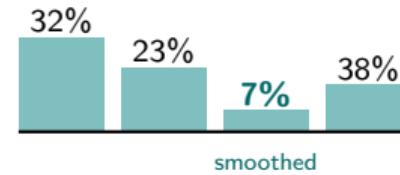
**Effect:**

- Empty bins get small but non-zero weight
- Popular bins still dominate

**Before: Empirical only**



**After: Laplace smoothed**



$$\hat{w}_{\text{lap}} = \frac{\text{count}+1}{\text{total}+\#\text{bins}}$$

# Agenda

## Part I: Paper Walkthrough

1. The identifiability problem
2. Functional ANOVA target
3. Mass-moving algorithm
4. Convergence & properties
5. Weight estimation ( $w$ )

## Part II: Experiments

### ► 6. German Credit (real data)

Main effects & interactions under different  $w$

### 7. Stress-test (synthetic)

How dependence  $\rho$  affects purification

## Conclusions

Summary & takeaways

# Our additions: two experiments

## Experiment 1 (Real data): German Credit

- Fit depth-2 XGBoost (piecewise-constant  $\Rightarrow$  tensorized effects)
- Purify same tensors under  $\hat{w}_{\text{emp}}$ ,  $\hat{w}_{\text{unif}}$ , and  $w_{\text{indep}}$
- Hypothesis: ignoring dependence in  $w$  reallocates mass between mains and interactions.

## Experiment 2 (Synthetic): stress-test under strong dependence

- Control dependence with  $\rho \in \{0, 0.3, 0.6, 0.9, 0.97\}$
- Compare purification under  $\hat{w}_{\text{emp}}$  vs misspecified  $w$  ( $\hat{w}_{\text{unif}}/w_{\text{indep}}/w_{\text{perm}}$ )
- Track divergence metrics and visualize surfaces
- Hypothesis: sensitivity to misspecified  $w$  increases with  $\rho$ .

# Experiment 1: German Credit (setup)

## ML hygiene

- Target: Credit amount (regression). Split: 75/25 train/test.
- Preprocessing: fill missing with unknown, one-hot encode categoricals.

## Model + purification

- Model: depth-2 XGBoost (`n_estimators=120, lr=0.1`).
- Binning: thresholds from tree splits (fixed across  $w$ ).
- Purification: compare  $\hat{w}_{\text{emp}}$ ,  $\hat{w}_{\text{unif}}$ , and  $w_{\text{indep}}$ .

## What to look for

Does the qualitative explanation (sign, monotonicity, quadrant pattern) change when dependence is ignored in  $w$ ?

# German Credit: Summary Statistics

Prediction invariance confirmed

RMSE = 2043.3 across all weight modes ( $\hat{w}_{\text{emp}}$ ,  $\hat{w}_{\text{unif}}$ ,  $w_{\text{indep}}$ ) — predictions are **identical**.

Top main effects (by variance)

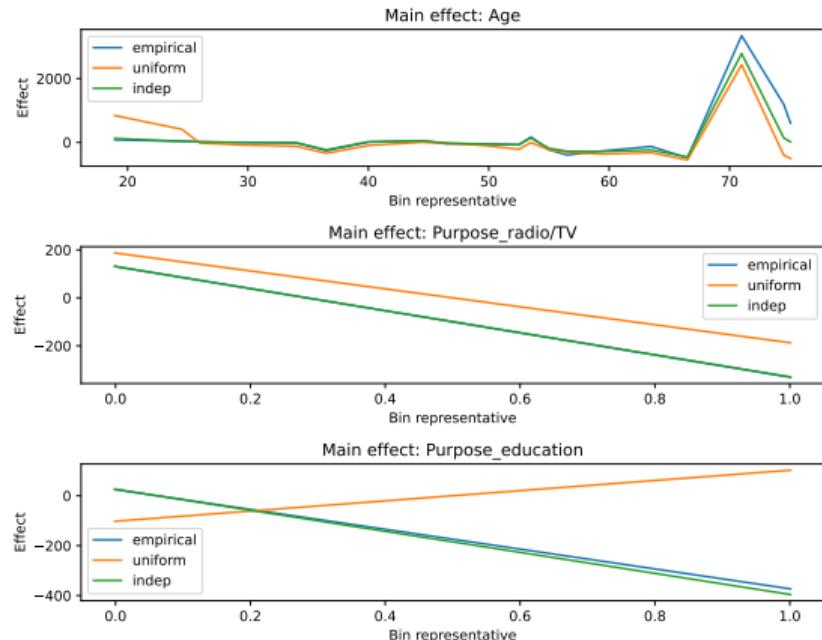
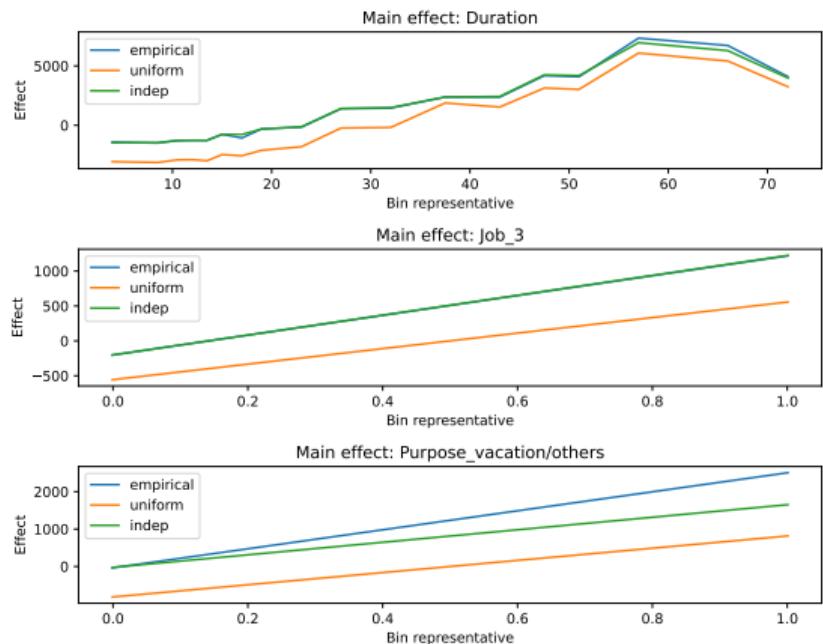
Feature	Var
Duration	2,906,619
Job_3	247,365
Purpose_vacation	93,542
Age	79,656
Purpose_radio/TV	43,529

Top interactions (by variance)

Pair	Var
Duration × Purpose_vac	29,522
Age × Duration	17,139
Duration × Job_3	7,498
Saving_mod × Purp_edu	5,888

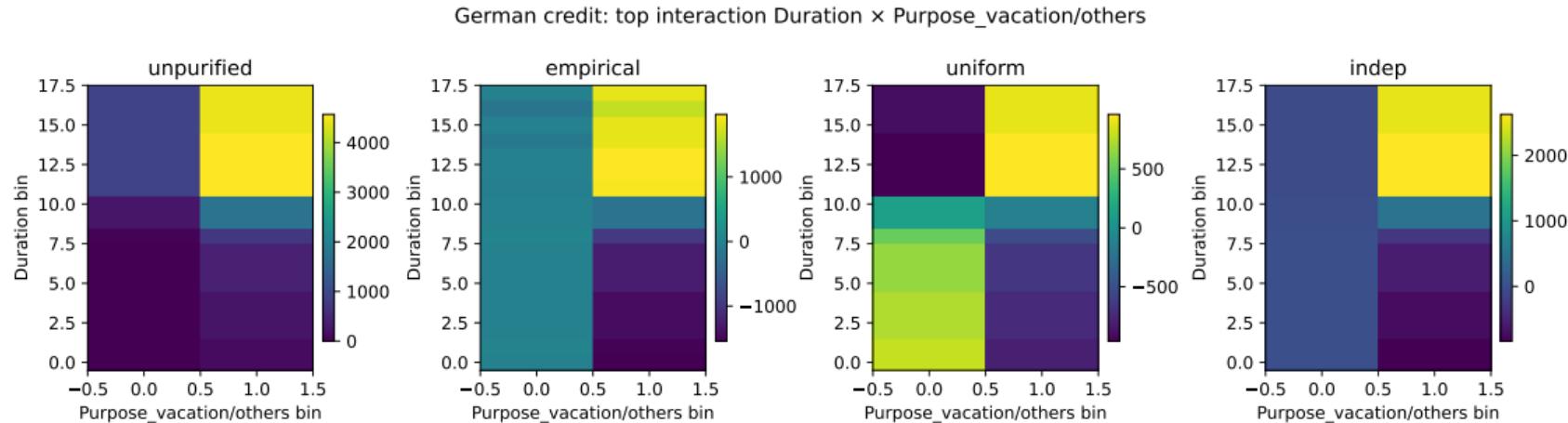
**Takeaway:** **Duration** dominates mains; its effect depends on loan purpose (interaction). Different  $w$  reallocates variance but predictions stay identical.

# German Credit: main effects (top features)



**Takeaway:** Same model and binning; differences across  $w$  are reallocations between mains and interactions.

# German Credit: top interaction heatmap



**Takeaway:** Interaction residual after purification. If  $w_{\text{indep}}$  changes the qualitative pattern relative to  $\hat{w}_{\text{emp}}$ , the interaction story is sensitive to ignoring dependence.

# Agenda

## Part I: Paper Walkthrough

1. The identifiability problem
2. Functional ANOVA target
3. Mass-moving algorithm
4. Convergence & properties
5. Weight estimation ( $w$ )

## Part II: Experiments

6. German Credit (real data)  
Main effects & interactions under different  $w$
- 7. **Stress-test (synthetic)**  
How dependence  $\rho$  affects purification

## Conclusions

Summary & takeaways

## Experiment 2: core question (weighting distribution $w$ )

### Question

Under strong feature dependence, how much do purified mains/interactions change when we purify using:

- the **correct joint** (empirical  $\hat{w}_{\text{emp}} \approx p(x)$ ) vs
  - a **misspecified** weighting (uniform or independence-assumed)?
- 
- Same learned piecewise-constant model  $\Rightarrow$  same tensors.
  - Only  $w$  changes.

# Stress-test: data generation

## Linear-Gaussian dependence

$$X_1 \sim \mathcal{N}(0, 1), \quad \varepsilon \sim \mathcal{N}(0, 1) \text{ independent}$$

$$X_2 = \rho X_1 + \sqrt{1 - \rho^2} \varepsilon, \quad \rho \in \{0.0, 0.3, 0.6, 0.9, 0.97\}$$

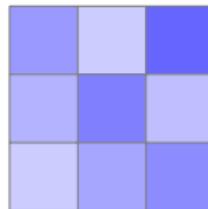
## Ground-truth functions

- (F1)  $F(X_1, X_2) = \beta_1 X_1 + \beta_2 X_2 + \gamma(X_1 X_2)$
- (F2)  $F(X_1, X_2) = \gamma \operatorname{sign}(X_1) \operatorname{sign}(X_2) + \beta_1 X_1 + \beta_2 X_2$

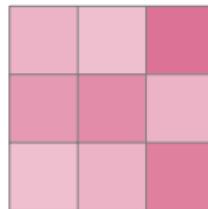
# Evaluation Metrics: Visual Guide

## Weighted $L_2$ Distance

$f_{12}^{\text{emp}}$  (reference)



$f_{12}^{\text{other}}$



$$d_{L_2} = \sqrt{\sum_{i,j} w_{ij} \left( f_{ij}^{\text{emp}} - f_{ij}^{\text{other}} \right)^2}$$

Measures overall surface difference

## Sign Flip Rate

$\text{Sign}(f^{\text{emp}})$  vs  $\text{Sign}(f^{\text{other}})$

✓	✓	✗
✗	✓	✓
✓	✗	✓

$$\text{flip rate} = \frac{\#\{\text{sign differs}\}}{\#\{\text{bins}\}} = \frac{3}{9}$$

Measures qualitative disagreement

**Takeaway:**  $L_2$  captures magnitude differences; sign flips capture qualitative interpretation changes (e.g., “positive effect” → “negative effect”).

# Stress-test: weighting comparison + metrics

We purify the same tensors under:

$$w \in \{\hat{w}_{\text{emp}}, \hat{w}_{\text{unif}}, w_{\text{indep}}, w_{\text{perm}}\}.$$

## Metrics vs $\rho$

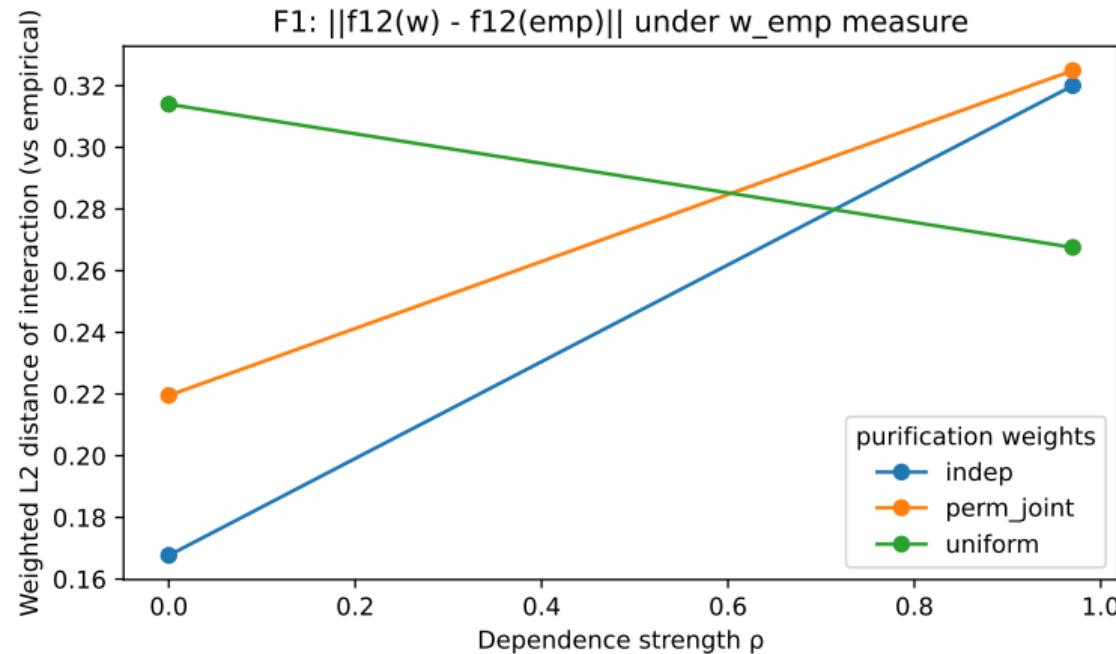
- Weighted  $L_2$  distance between purified interaction surfaces (reference measure:  $\hat{w}_{\text{emp}}$ )
- Sign flip rate vs  $\hat{w}_{\text{emp}}$  (fraction of bins with sign changes)
- Interaction variance share:  $\text{Var}_w(f_{12})/\text{Var}_w(F)$

# Weight Modes Explained

Mode	Definition
$\hat{w}_{\text{emp}}$ (empirical)	Observed joint distribution $\hat{p}(x_1, x_2)$ from training data. Captures dependence.
$\hat{w}_{\text{unif}}$ (uniform)	Equal weight $\frac{1}{ \Omega_1  \cdot  \Omega_2 }$ for all bin pairs. Ignores data distribution.
$w_{\text{indep}}$ (indep)	Product of marginals: $\hat{p}(x_1) \cdot \hat{p}(x_2)$ . Assumes independence analytically.
$w_{\text{perm}}$ (perm_joint)	Shuffle columns to break dependence, then compute empirical joint. Independence via resampling.

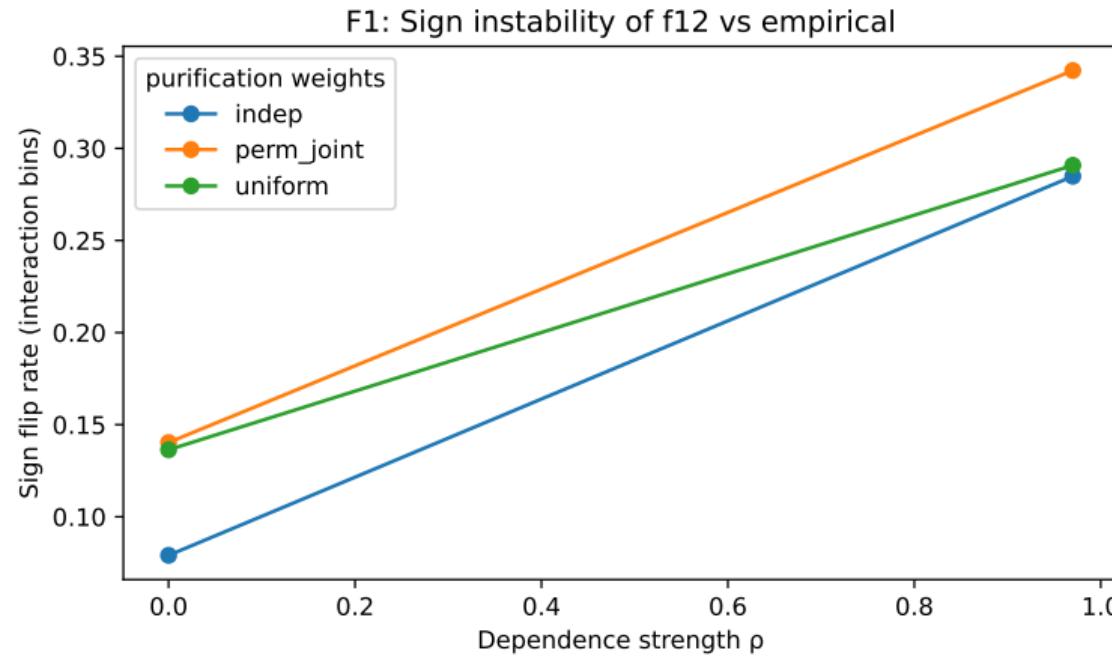
**Key point:** Under strong feature correlation, these weight choices lead to **different purified effects**.

# Stress-test (F1): $L_2$ distance vs $\rho$



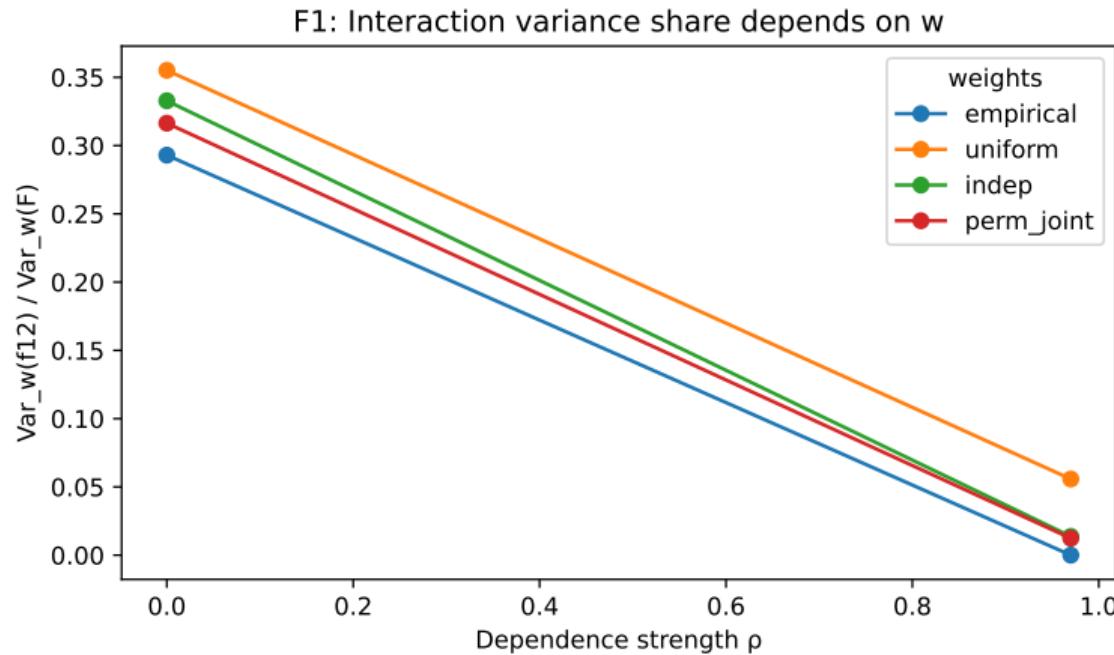
**Takeaway:** Distance from the  $\hat{w}_{\text{emp}}$ -purified interaction increases with dependence; misspecified  $w$  yields divergent explanations.

# Stress-test (F1): sign flip rate vs $\rho$



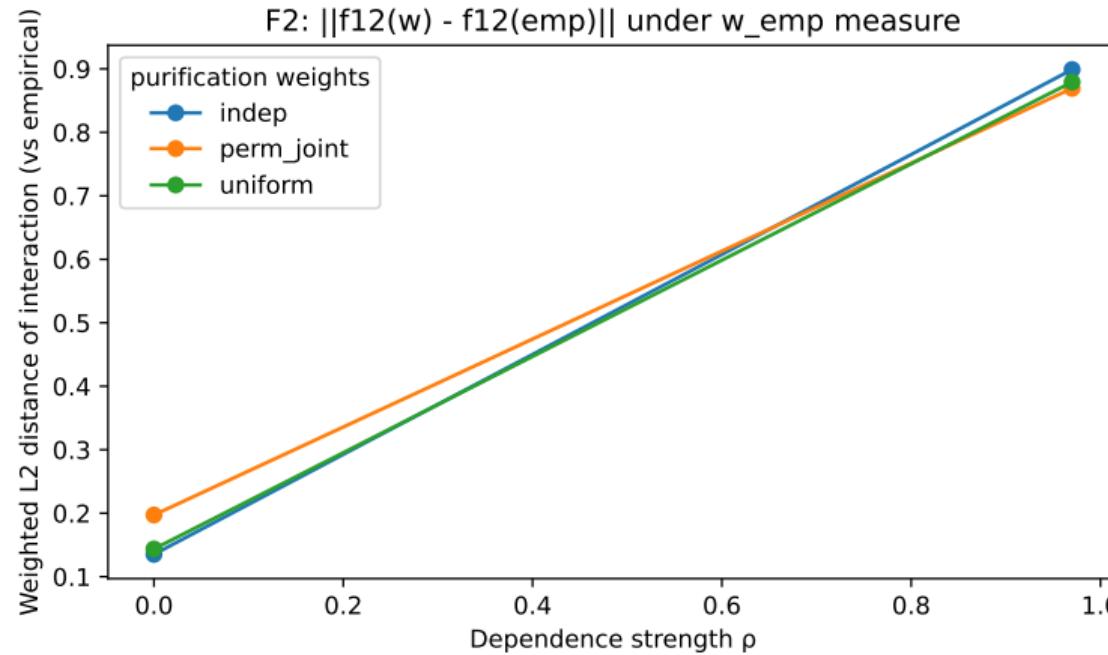
**Takeaway:** As dependence grows, qualitative interaction conclusions (sign) become sensitive to the choice of  $w$ .

# Stress-test (F1): interaction variance share vs $\rho$



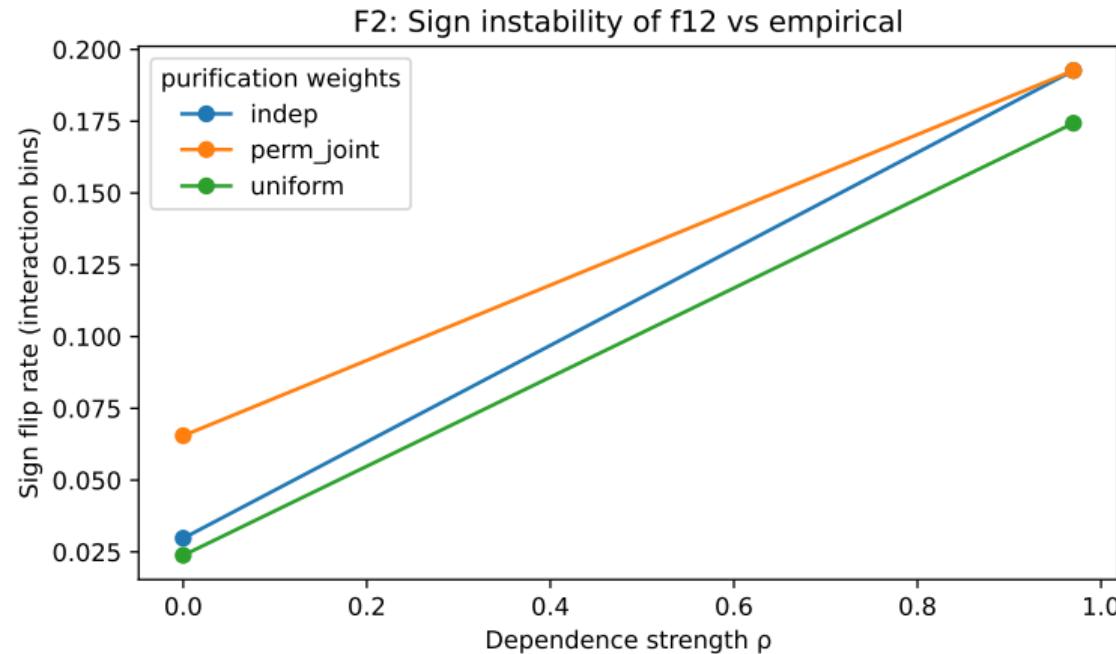
**Takeaway:** Variance attribution to the interaction changes with  $w$ ; under dependence, misspecifying  $w$  reshuffles main vs interaction credit.

## Stress-test (F2): $L_2$ distance vs $\rho$



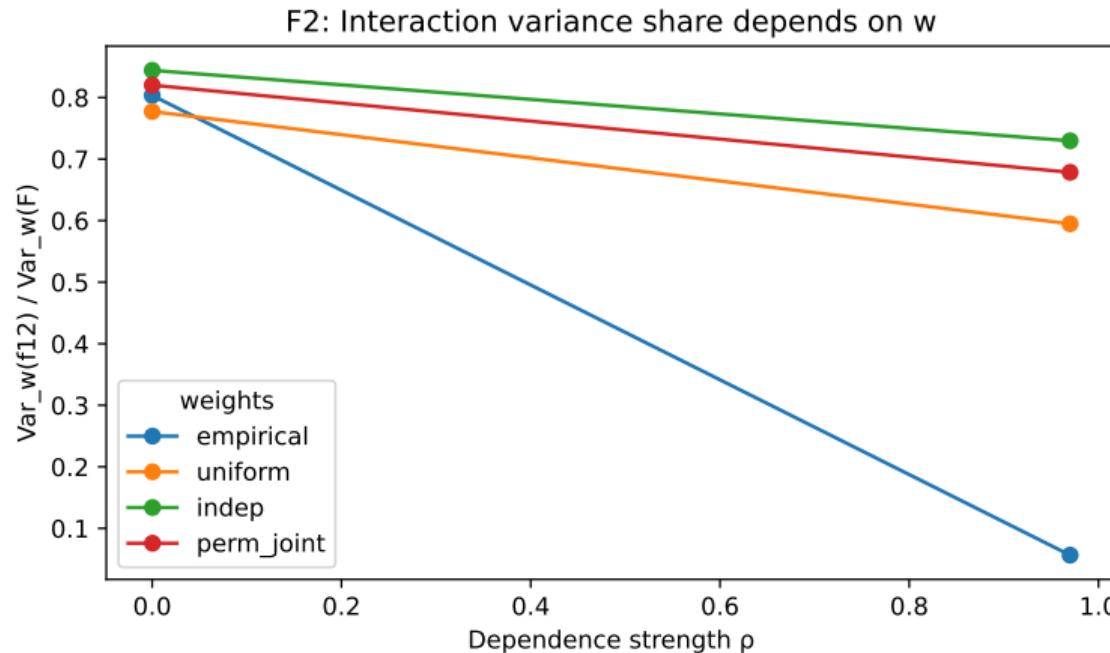
**Takeaway:** Distance from the  $\hat{w}_{\text{emp}}$ -purified interaction increases with dependence; misspecified  $w$  yields divergent explanations.

## Stress-test (F2): sign flip rate vs $\rho$



**Takeaway:** As dependence grows, qualitative interaction conclusions (sign) become sensitive to the choice of  $w$ .

## Stress-test (F2): interaction variance share vs $\rho$



**Takeaway:** Variance attribution to the interaction changes with  $w$ ; under dependence, misspecifying  $w$  reshuffles main vs interaction credit.

# Stress-test: Summary — $L_2$ Distance & Sign Flips

$L_2$  distance (interaction surface vs  $\hat{w}_{\text{emp}}$  reference)

Mode	F1 (multiplicative)		F2 (XOR-like)	
	$\rho=0$	$\rho=0.97$	$\rho=0$	$\rho=0.97$
uniform	0.31	0.27	0.14	<b>0.88</b>
indep	0.17	0.32	0.14	<b>0.90</b>
perm\_joint	0.22	0.32	0.20	<b>0.87</b>

Sign flip rate at  $\rho = 0.97$

Mode	F1	F2
uniform	29%	17%
indep	28%	19%

**Takeaway:** At high  $\rho$ , F2 shows  $L_2 \approx 0.9$  (surface almost unrecognizable). 20–30% of bins flip sign!

# Stress-test: Summary — Variance Attribution

Interaction variance share:  $\text{Var}_w(f_{12})/\text{Var}_w(F)$

Mode	F1		F2	
	$\rho=0$	$\rho=0.97$	$\rho=0$	$\rho=0.97$
$\hat{w}_{\text{emp}}$	29%	<b>0.002%</b>	80%	<b>5.6%</b>
uniform	36%	5.6%	78%	59%
indep	33%	1.4%	84%	<b>73%</b>

## Key insight

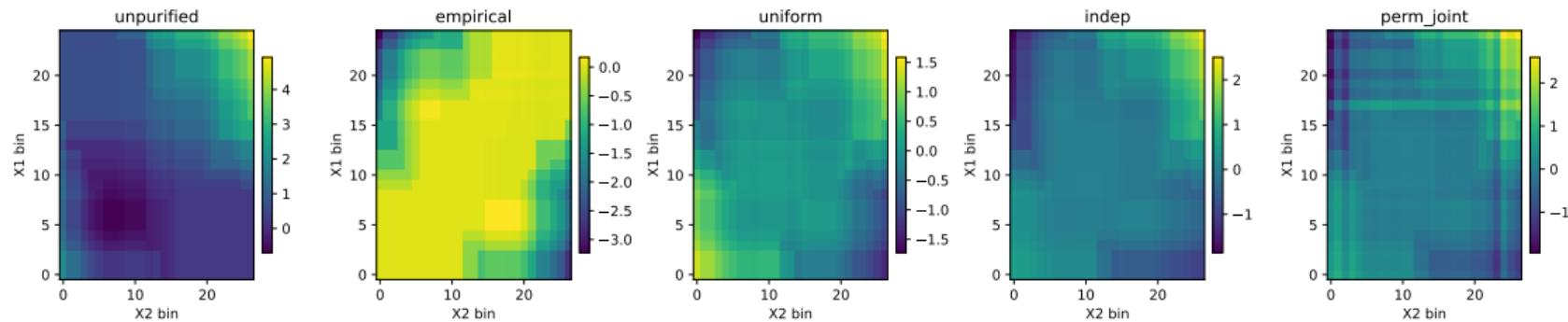
For F1 at  $\rho=0.97$ :  $\hat{w}_{\text{emp}}$  shows **0% interaction** (absorbed into mains).

But  $w_{\text{indep}}$  reports 1.4%,  $\hat{w}_{\text{unif}}$  reports 5.6% — **qualitatively different stories!**

**Takeaway:** Variance attribution shifts dramatically with  $w$ ; misspecifying  $w$  yields misleading importance rankings.

# Stress-test (F1): interaction heatmaps at $\rho = 0.97$

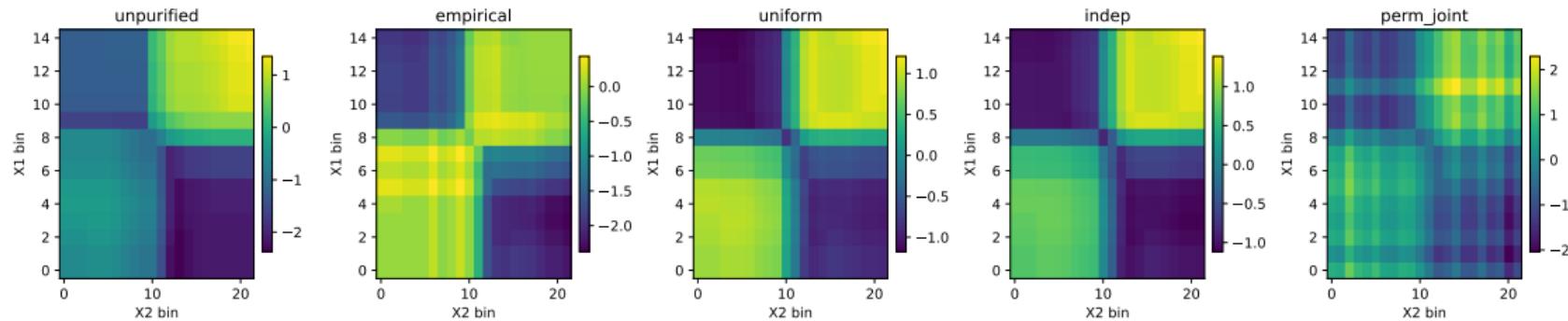
F1: interaction surface at  $\rho=0.97$  under different  $w$



**Takeaway:** At strong dependence ( $\rho = 0.97$ ), the interaction surface can change noticeably across  $w$ ;  $\hat{w}_{\text{emp}}$  emphasizes where samples occur.

# Stress-test (F2): interaction heatmaps at $\rho = 0.97$

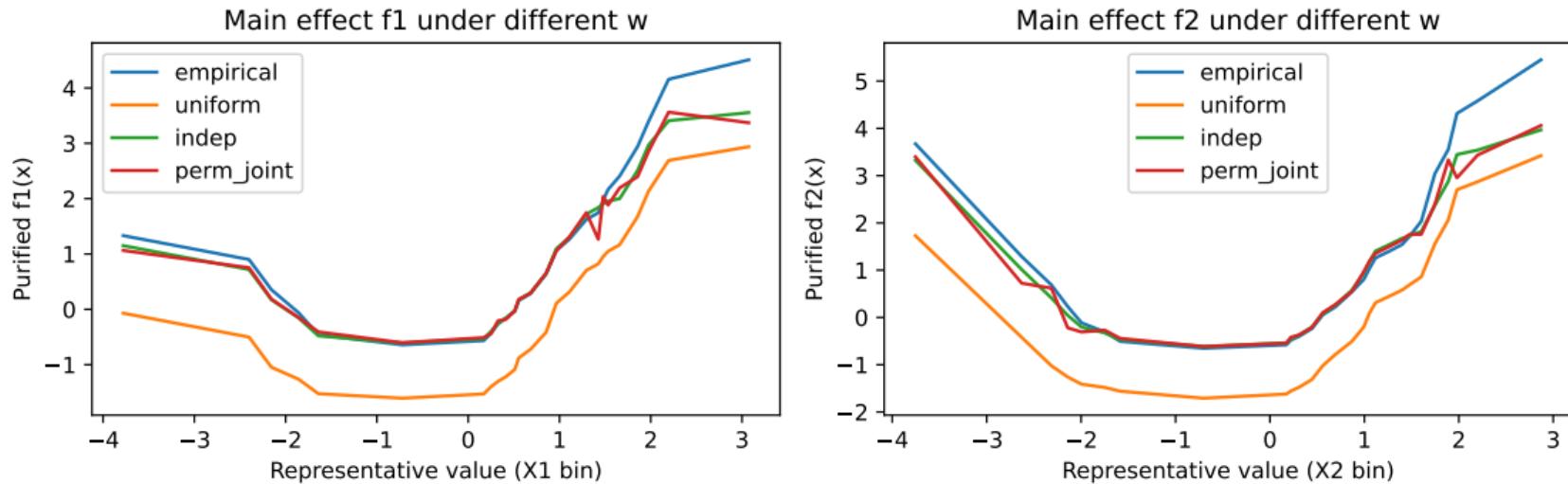
F2: interaction surface at  $\rho=0.97$  under different w



**Takeaway:** At strong dependence ( $\rho = 0.97$ ), the interaction story depends on  $w$ ; misspecification can change qualitative patterns.

# Stress-test: main effect curves at $\rho = 0.97$ (F1 example)

F1: main effects at  $\rho=0.97$



**Takeaway:** Main-effect curves can shift across  $w$  even when the model is fixed; changes are compensating reallocations with interactions (prediction stays the same).

# Conclusions

## Theoretical contribution (paper)

- Interactions are **not identifiable** without constraints; fANOVA yields a canonical decomposition under  $w$ .
- Purification: exact post-hoc mass-moving recovers fANOVA for piecewise-constant models.

## Our experimental findings

- **Predictions are invariant** to weight choice (confirmed: identical RMSE across  $w$ ).
- **Interpretations are NOT invariant:** under strong dependence ( $\rho=0.97$ ):
  - Variance attribution shifts dramatically (0% vs 73% interaction share)
  - 20–30% of bins flip sign
  - $L_2$  distance up to 0.9 (interaction surface almost unrecognizable)

**Takeaway:** Always report which  $w$  you used. Treat  $\hat{w}_{\text{emp}}$  vs  $w_{\text{indep}}$  as a **sensitivity analysis** when feature dependence is plausible.

# References & Resources

## Original Paper

Lengerich, B., Tan, S., Chang, C.-H., Hooker, G., & Caruana, R. (2020).

*Purifying Interaction Effects with the Functional ANOVA: An Efficient Algorithm for Recovering Identifiable Additive Models.*

AISTATS 2020, PMLR 108:2402–2412.

<https://proceedings.mlr.press/v108/lengerich20a.html>

## Code & Related Projects

- Original paper implementation: [https://github.com/blengerich/gam\\_purification](https://github.com/blengerich/gam_purification)
- InterpretML (Microsoft): <https://github.com/interpretml/interpret>

## Dataset

German Credit Data — UCI Machine Learning Repository

[https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data))

## AI Tools Used

The following AI tools were used to support development of this project:

- **NotebookLM** (Google) — understanding the paper
- **ChatGPT 5.2 Pro** (OpenAI) — code generation, debugging, and slide drafting
- **Claude Opus / Sonnet 4.5** (Anthropic) — code review, documentation, and slide refinement

All AI-generated content was reviewed, validated, and edited by the author.

# Thank You!

Questions?

[https://github.com/HaykTarkhanyan/fanova\\_purification](https://github.com/HaykTarkhanyan/fanova_purification)