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

Aik Tarkhanyan

Agenda

1. Problem: interactions are not identifiable
2. Definition: “pure” effects via weighted fANOVA
3. Algorithm: mass-moving purification for trees
4. Why w matters: interpretation depends on the chosen distribution
5. Our experiments: German Credit + dependence stress-test

Part I: Paper

Paper walkthrough

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:

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

- 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

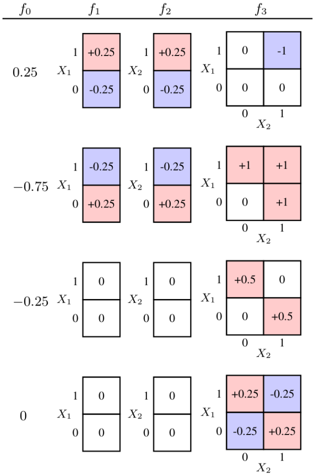
$$X_1 \vee X_2 = 0.25\,(X_1 \oplus X_2) + 0.5\,(X_1 - 0.5) + 0.5\,(X_2 - 0.5) + 0.75$$

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



f_0	f_1	f_2	f_3
0.25	$\begin{array}{cc} X_1 & \\ 1 & +0.25 \\ 0 & -0.25 \end{array}$	$\begin{array}{cc} X_2 & \\ 1 & +0.25 \\ 0 & -0.25 \end{array}$	$\begin{array}{ccc} X_1 & & \\ 1 & 0 & -1 \\ 0 & 0 & 0 \\ & 0 & 1 \\ & X_2 & \end{array}$
-0.75	$\begin{array}{cc} X_1 & \\ 1 & -0.25 \\ 0 & +0.25 \end{array}$	$\begin{array}{cc} X_2 & \\ 1 & -0.25 \\ 0 & +0.25 \end{array}$	$\begin{array}{ccc} X_1 & & \\ 1 & +1 & +1 \\ 0 & 0 & +1 \\ & 0 & 1 \\ & X_2 & \end{array}$
-0.25	$\begin{array}{cc} X_1 & \\ 1 & 0 \\ 0 & 0 \end{array}$	$\begin{array}{cc} X_2 & \\ 1 & 0 \\ 0 & 0 \end{array}$	$\begin{array}{ccc} X_1 & & \\ 1 & +0.5 & 0 \\ 0 & 0 & +0.5 \\ & 0 & 1 \\ & X_2 & \end{array}$
0	$\begin{array}{cc} X_1 & \\ 1 & 0 \\ 0 & 0 \end{array}$	$\begin{array}{cc} X_2 & \\ 1 & 0 \\ 0 & 0 \end{array}$	$\begin{array}{ccc} X_1 & & \\ 1 & +0.25 & -0.25 \\ 0 & -0.25 & +0.25 \\ & 0 & 1 \\ & X_2 & \end{array}$

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

4) Multiplicative model and the (α, β) degree of freedom

A classic interaction model:

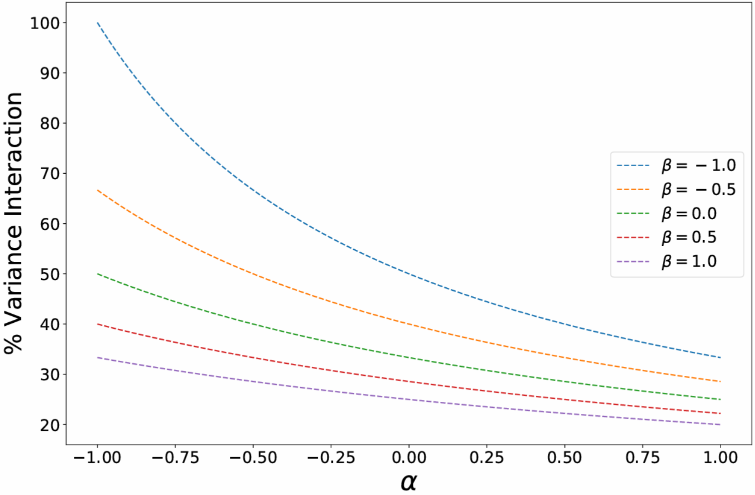
$$Y \approx a + bX_1 + cX_2 + dX_1X_2.$$

An equivalent re-parameterization (for any α, β):

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

- Identifiable as a function, but coefficients are **not meaningful** without rules for choosing α, β.
- 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 (α, β)



Takeaway: As (α, β) vary, variance attribution shifts between mains and interactions even though predictions do not.

5) Functional ANOVA (fANOVA): the canonical target

The functional ANOVA decomposes a function into orthogonal components under a weight $w(X)$:

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

For continuous F, the weighted fANOVA solves a least-squares projection under w with orthogonality constraints. A convenient equivalent condition is **integrate-to-zero** for each slice:

$$\forall u,\ \forall i \in u : \quad \int f_u(X_u)\, w(X)\, dX_i\, dX_{-u} = 0.$$

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

For piecewise-constant F on bins Ω_j (one set of bins per feature), each f_u becomes a tensor of effect sizes. The integrate-to-zero constraints become weighted **slice mean-zero** constraints on the tensor:

$$\forall u, \forall i \in u, \forall X_{u \setminus i} : \sum_{x_i \in \Omega_i} f_u(X_{u \setminus i}, X_i = x_i) \sum_{X_{-u}} w(X) = 0.$$

- Pairwise intuition: for f_{ij}, each row/column has weighted mean zero (conditioning removes lower-order structure).
- This is the key observation enabling an exact post-hoc algorithm for trees.

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

Define **pure interactions** as variance that cannot be represented by any smaller subset of variables.

Definition (paraphrased)

Find functions $\{f_u\}$ minimizing squared error to $\mathbb{E}[Y \mid X]$ under the data distribution, subject to $\mathbb{E}[f_u(X_u) \mid X_v] = 0$ for all strict subsets $v \subset u$.

Crucially, this is equivalent to the fANOVA decomposition with $w(X) = p(X)$.

8) Purification for trees: mass-moving

For tree-based models, we can represent effects as tensors T_u. Idea: move the weighted mean of each 1D slice from a higher-order tensor into its lower-order tensor, without changing predictions.

Weighted slice mean

$m(T_u, i, X_{u\setminus i})$ = weighted mean of the slice along i at fixed $X_{u\setminus i}$.

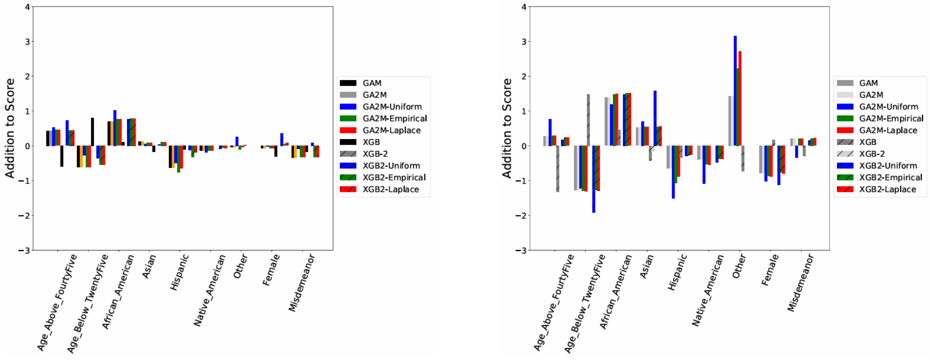
- **Purify-Matrix**: iteratively zero slice means of T_u (push mass into $T_{u\setminus i}$).
- **Purify**: run from highest-order to lowest-order tensors.

Mass-moving (Purify-Matrix): 3-step schematic

1. Compute the weighted slice mean of the interaction tensor along one dimension.
2. Subtract that mean from the interaction slice (enforcing slice mean ≈ 0).
3. Add the same amount to the corresponding lower-order term (main effect / intercept).

Takeaway: Each sweep pushes lower-order structure down while keeping predictions unchanged (Algorithm 1: Purify-Matrix).

Example: COMPAS



(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.

Example: California housing

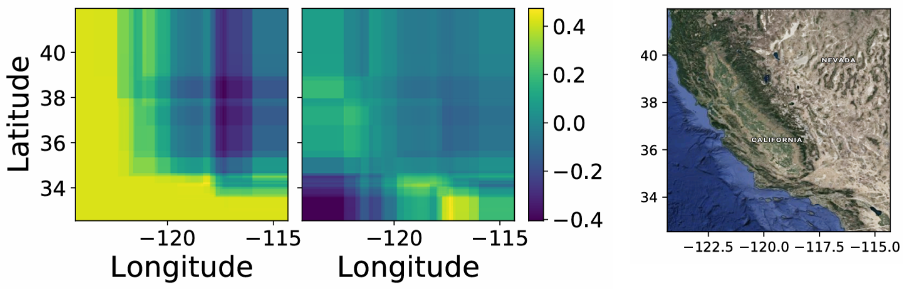


Figure 5: Interaction of the Latitude/Longitude features in an XGB2 model trained on the California housing data. The left pane is the unpurified interaction, the middle pane is the purified interaction, and the right is the map of California from which samples were drawn. Purification sorts out the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

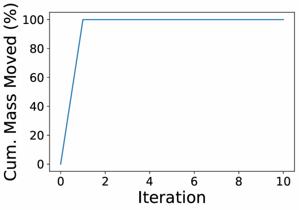
For many simple weights (e.g., uniform along row/column dimensions), purification converges in a single pass.

Theorem 2 (generic non-degenerate w)

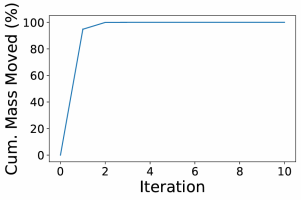
Converges to tolerance ε in $O(\log(M_0) - \log(\varepsilon))$ iterations per interaction tensor.

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

Part II: Experiments

Experiment 1: German Credit

Our additions: two experiments

Experiment 1 (Real data): German Credit

- Fit depth-2 XGBoost (piecewise-constant ⇒ tensorized effects)
- Use tree split thresholds as bins
- Purify the same tensors under $\hat{w}_{\text{emp}}$, $\hat{w}_{\text{unif}}$, and a deliberate misspecification w_{indep}
- Hypothesis: ignoring dependence in w reallocates mass between mains and interactions.

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

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

Experiment 1: German Credit (setup)

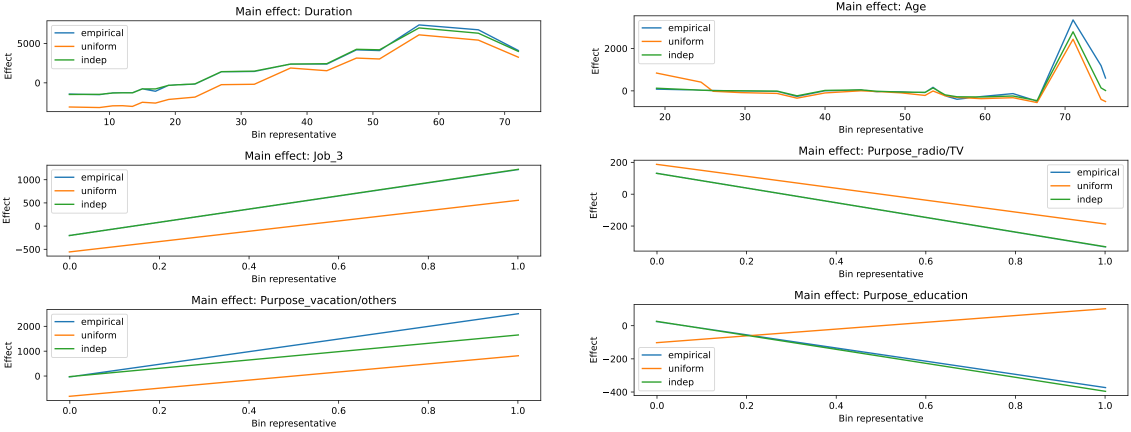
ML hygiene

- Target (this repo's CSV): `Credit amount` (regression). If a label exists (e.g., `Risk`), treat it as binary good/bad classification.
- Split: train/test via `train_test_split` with test_size = 0.25 (random_state = 0).
- Metric: RMSE on the test set (classification case: ROC-AUC on predicted probabilities; objective is `binary:logistic`).
- Preprocessing: fill missing categoricals with `unknown`, one-hot encode categoricals (`get_dummies`), cast to float; no scaling.

Model + purification

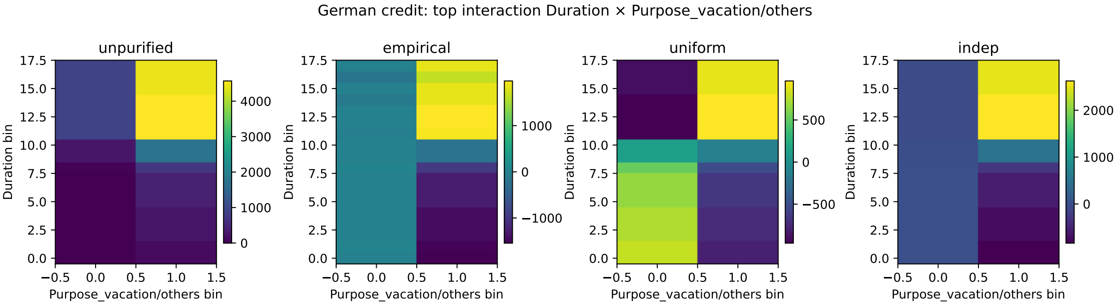
- Model: depth-2 XGBoost (n_estimators = 120, learning_rate = 0.1).
- Binning: thresholds derived from the learned tree splits (fixed across w).
- Purification: compare decompositions under $\hat{w}_{\mathrm{emp}}$, $\hat{w}_{\mathrm{unif}}$, and w_{indep}.

German Credit: main effects (top features)



Takeaway: Same model and binning; differences across w are reallocations between mains and interactions. Compare $\hat{w}_{\mathrm{emp}}$ vs w_{indep} for dependence sensitivity.

German Credit: top interaction heatmap



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

Experiment 2: Stress-test under strong dependence

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 $\hat{w}_{\mathrm{emp}} \approx p(x)$) 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) $F(X_1, X_2) = \gamma\,\mathrm{sign}(X_1)\mathrm{sign}(X_2) + \beta_1 X_1 + \beta_2 X_2$

Stress-test: weighting ablation + metrics

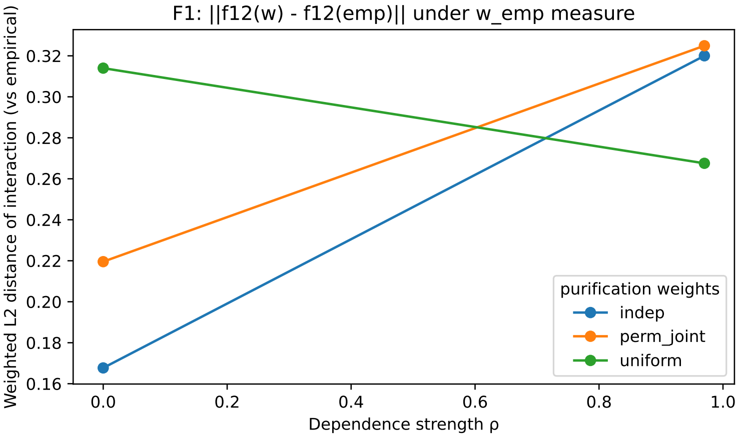
We purify the same tensors under:

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

Metrics vs ρ

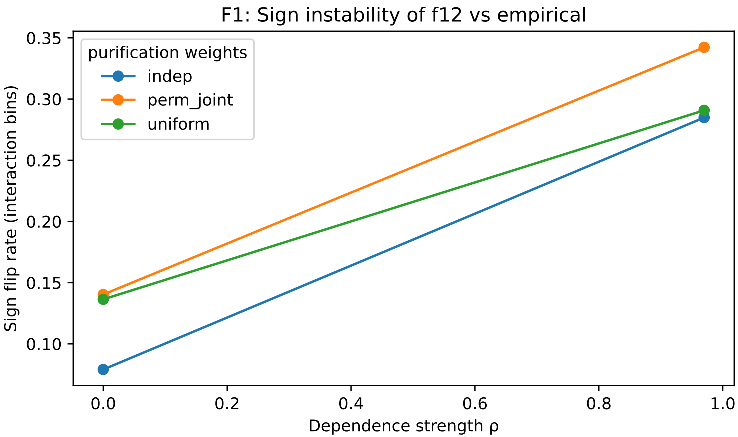
- Weighted L_2 distance between purified interaction surfaces (reference measure: $\hat{w}_{\mathrm{emp}}$)
- Sign flip rate vs $\hat{w}_{\mathrm{emp}}$ (fraction of bins with sign changes)
- Interaction variance share: $\mathrm{Var}_w(f_{12})/\mathrm{Var}_w(F)$

Stress-test (F1): L_2 distance vs ρ



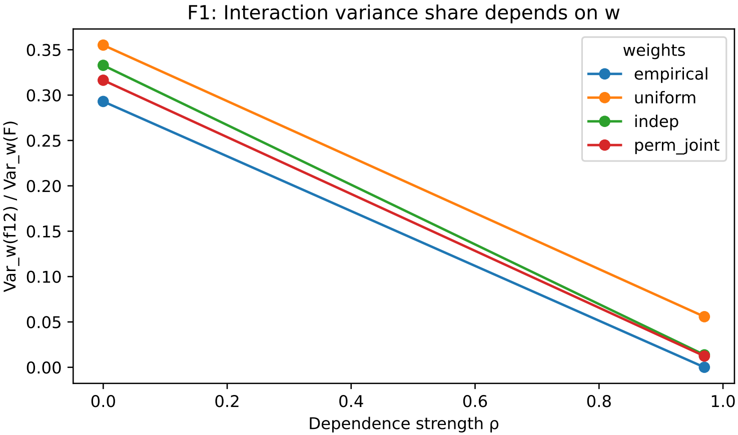
Takeaway: Distance from the $\hat{w}_{\mathrm{emp}}$-purified interaction increases with dependence; misspecified w yields divergent explanations.

Stress-test (F1): sign flip rate vs ρ



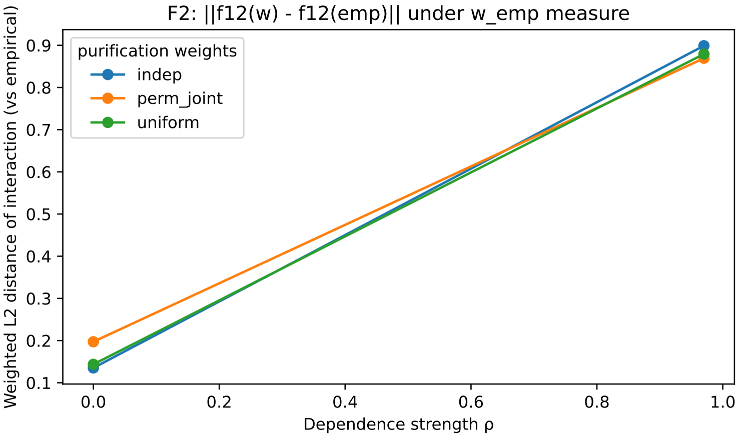
Takeaway: As dependence grows, qualitative interaction conclusions (sign) become sensitive to the choice of w.

Stress-test (F1): interaction variance share vs ρ



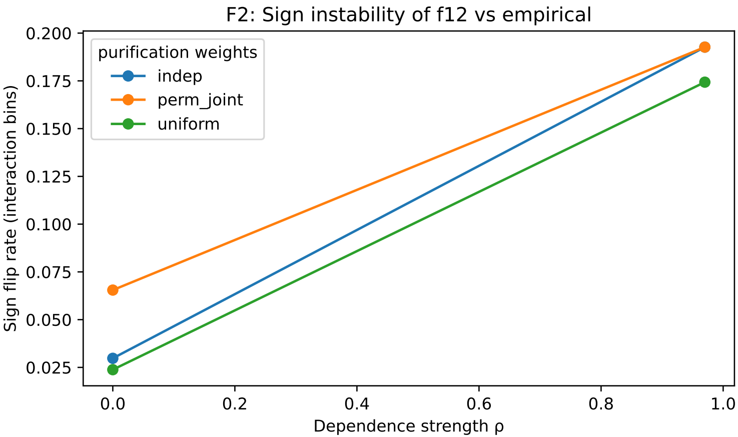
Takeaway: Variance attribution to the interaction changes with *w*; under dependence, misspecifying *w* reshuffles main vs interaction credit.

Stress-test (F2): L_2 distance vs ρ



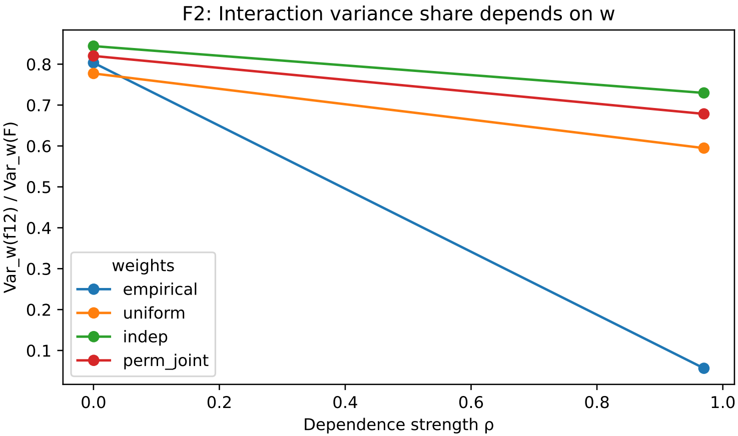
Takeaway: Distance from the $\hat{w}_{\mathrm{emp}}$-purified interaction increases with dependence; misspecified w yields divergent explanations.

Stress-test (F2): sign flip rate vs ρ



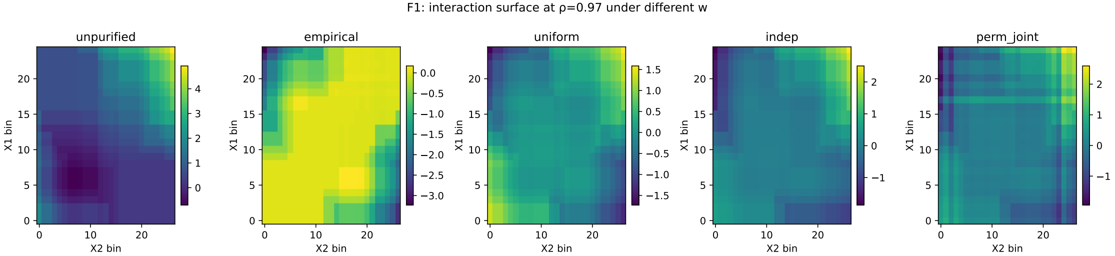
Takeaway: As dependence grows, qualitative interaction conclusions (sign) become sensitive to the choice of w.

Stress-test (F2): interaction variance share vs ρ



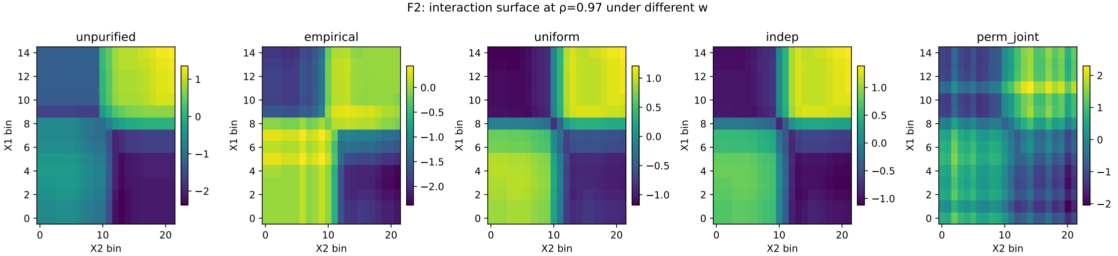
Takeaway: Variance attribution to the interaction changes with *w*; under dependence, misspecifying *w* reshuffles main vs interaction credit.

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



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 ρ=0.97 under different w

unpurified

empirical

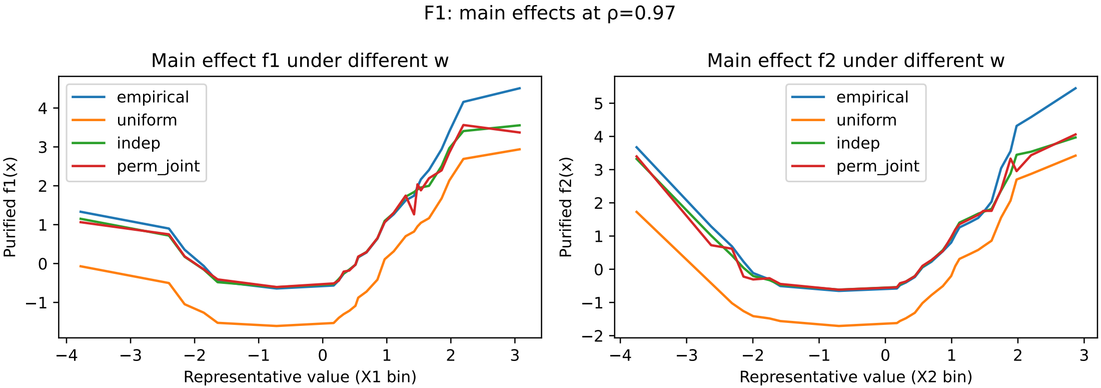
uniform

indep

perm_joint

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)



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

- Interactions are not identifiable without constraints; fANOVA yields a canonical decomposition under w.
- Purification: exact post-hoc mass-moving recovers fANOVA for piecewise-constant (tree) models.
- Our experiments: under dependence, misspecifying w can materially change interpretations.
- Practical: always report which w you used; treat $\hat{w}_{\mathrm{emp}}$ vs w_{indep} as a sensitivity analysis when dependence is plausible.