SEPARATING COVARIATE SHIFT FROM MECHANISM CHANGE WITH TWO DISCRIMINATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

When labeled datasets differ, deciding what to inspect next requires separating a change in input frequencies from a change in the labeling rule. We develop a two-component diagnostic based on a balanced dataset source indicator Z: input information $I(Z;X)$ and conditional information $I(Z;Y \mid X)$, called conditional Jensen–Shannon discrepancy (CJSD). The components add to the joint Jensen–Shannon divergence. CJSD isolates label-rule differences on shared support, while the input component reveals how much conditional information remains observable. Two source classifiers estimate both components through held-out log losses, whose record-level differences connect a global score to regions worth inspecting. Controlled dataset pairs illustrate the separation and localize an imposed change in electricity records. Unaltered gas-sensor measurements demonstrate why a small conditional score can coexist with poor transfer when the inputs strongly separate. An equal-observation comparison with established conditional-information estimators assesses magnitude recovery, null behavior, and rotation sensitivity. Together, these results provide a practical way to read dataset differences and the limits of the evidence they supply.

1 INTRODUCTION

Two labeled datasets can differ in two distinct ways: the inputs that occur can change, or the labels associated with the same inputs can change. The distinction guides what to inspect. Reweighting addresses changed input frequencies when the labeling rule is stable and coverage is adequate; a changed rule instead calls for examining labels and predictor validity. An input-only distance cannot make this distinction, and a drop in prediction accuracy does not say which part changed.

We study this question through two ordinary prediction tasks. Pool the datasets with equal source probabilities, and tag each observation by its source Z. First predict Z from X. Then also supply Y. The improvement from seeing the label measures source information that the inputs did not already provide. The Shannon chain rule gives

$$I(Z;X,Y) = \underbrace{I(Z;X)}_{\text{input component } I_x} + \underbrace{I(Z;Y \mid X)}_{\text{conditional component } D_{\mathrm{CJS}}} . \tag{1}$$

With balanced sources, the left side is the joint Jensen–Shannon divergence and the first term is its input counterpart (Lin, 1991; Cover & Thomas, 2006). We call the conditional term *conditional Jensen–Shannon discrepancy* (CJSD), reflecting its representation as an overlap-weighted average of conditional Jensen–Shannon divergences. It is ordinary conditional mutual information (CMI), applied here to the dataset source.

Figure 1 gives the intended reading. A conditional component of zero is compatible both with a shared labeling rule and with completely separated input supports. The input component distinguishes these situations: it also measures how much source uncertainty remains for labels to resolve. Thus the two numbers belong together.

We use *covariate shift* for changed $P(X)$ with fixed $P(Y \mid X)$ (Shimodaira, 2000), and *mechanism change*, or *label-rule change*, for a changed statistical conditional on shared support (Webb et al., 2016; Gama et al., 2014). This use of "mechanism" does not assert causation. Both datasets must have labels with common meanings.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

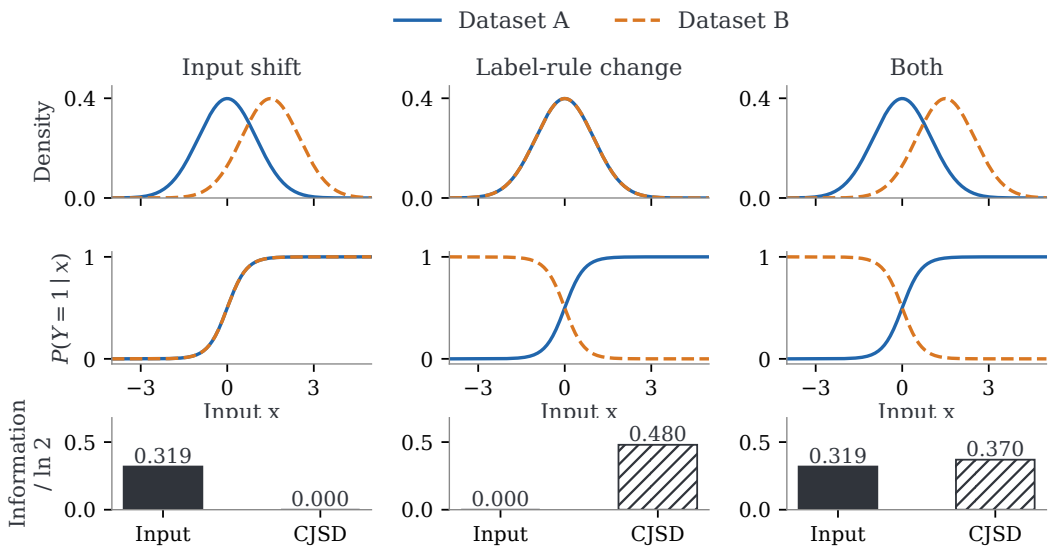

Figure 1: One comparison, two questions. Each column shows input densities, label probabilities, and the two population components in bits. Input shift changes which inputs occur; label-rule change alters labels at a given input. The same rule reversal has a smaller conditional component when the inputs also separate (right). Bars use numerical integration of the displayed Gaussian/logistic laws, specified in Appendix C.

Relation to CMI estimation. CCMI already estimates CMI using classifiers (Mukherjee et al., 2019); it can estimate the same target. Our focus is the diagnostic reading of two dataset components and a direct source-classification implementation. Because Z is an observed binary label, ordinary probabilistic classifiers estimate $H(Z \mid X)$ and $H(Z \mid X, Y)$ from log loss, without creating the joint-versus-product classification problems used by CCMI's MI-difference route. The distinct objectives are compared under matched neural budgets in Section 5. The contribution is this diagnostic formulation, implementation, and validation, rather than a new chain rule or a new information quantity.

Contributions.

1. **A two-component reading of dataset change.** We connect the additive decomposition to conditional equality on shared support, observable overlap, and a changed-label fraction for deterministic labels with equal input laws (Section 2).
2. **A direct estimator with inspectable contributions.** Two source classifiers give both components from paired held-out losses. Their signed excess-risk identity explains inflation and attenuation; individual loss differences connect the conditional score to input regions (Section 3).
3. **Evidence organized around how to read the result.** Controlled changes show separation and localization; unaltered sensor records illustrate limited overlap in practice. A common-target comparison with CCMI and a mixed-variable nearest-neighbor estimator reports recovery and failures across dimensions and rotations (Sections 4–5).

2 What the two components mean

Let P_A and P_B be distributions on common inputs X and a finite label set $\mathcal{Y}$. Write their regular conditional label distributions as $\eta_A(\cdot \mid x)$ and $\eta_B(\cdot \mid x)$ (Kallenberg, 2021). Draw $Z \sim \mathrm{Bern}(1/2)$, then draw (X, Y) from the corresponding dataset law. Write $\tilde{\mu} = (P_A^X + P_B^X)/2$ and $w(x) = P(Z = 0 \mid X = x)$. All theoretical expressions use natural logarithms; dividing by $\ln 2$ expresses the components in bits.

Lemma 1 (Conditional representation). *Let H denote Shannon entropy and $\mathrm{JS}_w(p,q) = H(wp + (1-w)q) - wH(p) - (1-w)H(q)$. Then*

$$I_x = \mathrm{JS}_{1/2}(P_A^X, P_B^X), \tag{2}$$

$$D_{\mathrm{CJS}} = I(Z;Y \mid X) = \mathbb{E}_{\tilde{\mu}}\mathrm{JS}_{w(X)}(\eta_A(\cdot \mid X), \eta_B(\cdot \mid X)), \tag{3}$$

and $I_x + D_{\mathrm{CJS}} = \mathrm{JS}_{1/2}(P_A, P_B)$. In particular,

$$0 \le D_{\mathrm{CJS}} \le H(Z \mid X) = \ln 2 - I_x. \tag{4}$$

A conditional null that allows input shift. The conditional component vanishes under arbitrary changes to the input law if the conditional label law stays the same. More precisely:

Proposition 1 (Equality on shared support). *$D_{\mathrm{CJS}} = 0$ if and only if $\eta_A = \eta_B$ at $\tilde{\mu}$-almost every x with $0 < w(x) < 1$. If the input laws are mutually singular, then $D_{\mathrm{CJS}} = 0$ for every pair of conditional label laws.*

The overlap condition is part of the quantity's meaning. If inputs alone identify the source, labels cannot add source information. Equation (4) therefore makes a large input component a reason to inspect coverage before reading a small conditional component as agreement. This is an observability limit, not a claim that prediction transfers across nonoverlapping regions.

A direct interpretation for deterministic labels. Let $h(u) = -u\ln u - (1-u)\ln(1-u)$, with $0\ln 0 = 0$.

Proposition 2 (Overlap-weighted disagreement). *If $Y = f_T(X)$ in dataset $T \in \{A, B\}$, then*

$$D_{\mathrm{CJS}} = \mathbb{E}_{\tilde{\mu}}\left[h(w(X))\,\mathbf{1}\{f_A(X) \neq f_B(X)\}\right].$$

For equal input laws $P_A^X = P_B^X = P_X$, this becomes $D_{\mathrm{CJS}}/\ln 2 = P_X(f_A(X) \neq f_B(X))$.

With equal input laws, normalized CJSD is exactly the fraction of inputs whose deterministic label changes. With unequal laws, the entropy weight downweights regions where source identity is already apparent from inputs. These elementary consequences of the information identities specify what the diagnostic measures. Appendix A gives proofs. Pair-dependent overlap also means that CJSD is a discrepancy, not a metric.

3 Two classifiers and a paired loss difference

Fit and evaluate. Balance the sources by sampling $n = \min(n_A, n_B)$ rows from each. Use three source-stratified folds shared by an input-only classifier $T_1 : X \to Z$ and a joint classifier $T_2 : (X, Y) \to Z$. In each fold, fit both on the other folds and evaluate on held-out rows. Encode categorical Y by indicator columns. Clip predicted source probabilities to $[10^{-4}, 1 - 10^{-4}]$ before computing log loss. For $N = 2n$ pooled rows and held-out losses $\ell_i^{(1)}, \ell_i^{(2)}$,

$$\widehat{I_x} = \ln 2 - \frac{1}{N}\sum_i \ell_i^{(1)}, \qquad \widehat{D_{\mathrm{CJS}}} = \frac{1}{N}\sum_i \left(\ell_i^{(1)} - \ell_i^{(2)}\right). \tag{5}$$

The estimated components add exactly to $\ln 2 - \widehat{R}_2$. We retain signed estimates: negative fitted values expose estimation error even though population information is nonnegative. Classifier settings and grouping of duplicate observations are specified with each experiment in the appendix.

Connect a score to records. Every held-out row contributes $c_i = (\ell_i^{(1)} - \ell_i^{(2)})/\ln 2$ to the normalized conditional score. Averaging within an input region shows where labels improve source prediction; count-weighted region means recover the global score. This is a descriptive decomposition of the fitted result. Individual contributions can be negative, and region averages are not calibrated local tests or estimates of a causal effect.

Separate learning error from evaluation noise. For a fixed fitted pair, let R_j denote fresh-data expected log loss, $U_1 = X$, and $U_2 = (X, Y)$. The log-loss identity gives

$$\widetilde{D} := R_1 - R_2 = D_{\mathrm{CJS}} + \epsilon_1 - \epsilon_2, \qquad \epsilon_j = \mathbb{E}\mathrm{KL}(P(Z \mid U_j) \,\|\, T_j(\cdot \mid U_j)) \geq 0, \tag{6}$$

where KL is Kullback–Leibler divergence (Cover & Thomas, 2006) and clipping is included in T_j. Input-only excess risk inflates the gap; joint excess risk attenuates it. Richer joint inputs alone do not order the risks. A small fitted score can reflect a small target, limited overlap, joint-classifier error, or evaluation noise. This identity interprets learner sensitivity; it does not identify the unknown excess risks.

The identity also holds for averaged foldwise population gaps. Reported across-seed standard deviations (SDs) describe repeated fitting and sampling, not confidence intervals for Shannon CJSD: these also require learning-error control. Appendix A states consistency conditions.

The input-only loss has a further useful interpretation: for any fixed probabilistic T_1,

$$D_{\mathrm{CJS}} \leq H(Z \mid X) \leq R_1. \tag{7}$$

Thus a small population source-prediction loss limits how much conditional information can be observed, without requiring an optimal classifier. Replacing R_1 by held-out loss still requires evaluation-uncertainty control before this becomes a confidence bound.

Choosing classifiers for a new comparison. Compare candidates by each classifier's validation log loss separately, within the outer training folds; reserve outer held-out rows for reporting. Select for source prediction, not for the largest CJSD gap. Report both losses against the balanced constant predictor's $\ln 2$: larger losses flag avoidable predictive error, while equality is compatible with a true null. Lower individual risks need not improve their difference because errors can cancel. Our comparisons below use fixed settings without tuning.

4 Reading Changes from Observations

All experimental scores are divided by $\ln 2$. We organize the evidence around three questions: do the components distinguish known changes, where does a conditional signal arise, and how should a small signal be read when the inputs strongly separate?

4.1 Do the components separate known kinds of change?

The constructed battery comprises 130 synthetic pairs and 36 pairs each from Electricity (Harries, 1999) and Covertype (Blackard, 1998). Each pair has 7,000 observations per source. Synthetic laws have Gaussian inputs and nonlinear Bernoulli label rules. Real-record constructions use input-dependent sampling and regional label remapping. Their change types are specified interventions, rather than inferred historical events. Three-fold random-forest (RF) classifiers (Breiman, 2001) estimate both components. Appendix B specifies all conditions and seeds.

Figure 2 shows the distinction directly. Within each family, the conditional scores rank all label-rule-change pairs above all covariate-only pairs: area under the receiver operating characteristic curve (AUC) is 1.00. This compares 116 label-rule pairs with 42 covariate-only pairs and excludes null and combined conditions. It validates separation in this battery; it does not establish accurate magnitudes or general detection power. Section 5 therefore uses known numerical targets, small changes, and two null controls.

4.2 Where does the conditional score come from?

Electricity contains half-hourly market records from New South Wales (NSW) and Victoria. The UP/DOWN label compares the NSW price with its preceding 24-hour moving average (Harries, 1999). In Figure 3, inputs are unchanged and 2,096 of 7,000 source-B labels are reversed above a demand threshold. The visual sequence starts with the same records before and after this imposed remapping, then shows their held-out contributions, then the global result.

The conditional contribution rises inside the remapped region; the original-label control stays near zero. Regional means of 0.347 inside and -0.0025 outside, with pooled region mass 0.3045, average

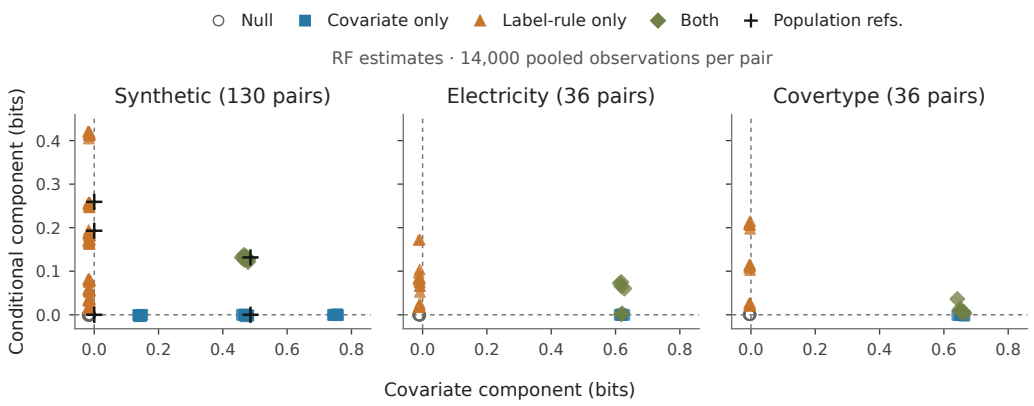

Figure 2: The two-component reading across all 202 constructed pairs. Each point is one condition–seed pair, estimated with RF classifiers. Covariate-only changes move along the input axis; label-rule changes move along the conditional axis; both can occur together. All signed estimates are shown, without jitter; coincident points overlap. Black crosses are five synthetic population references from Appendix C.1. Dashed lines mark zero, not test thresholds.

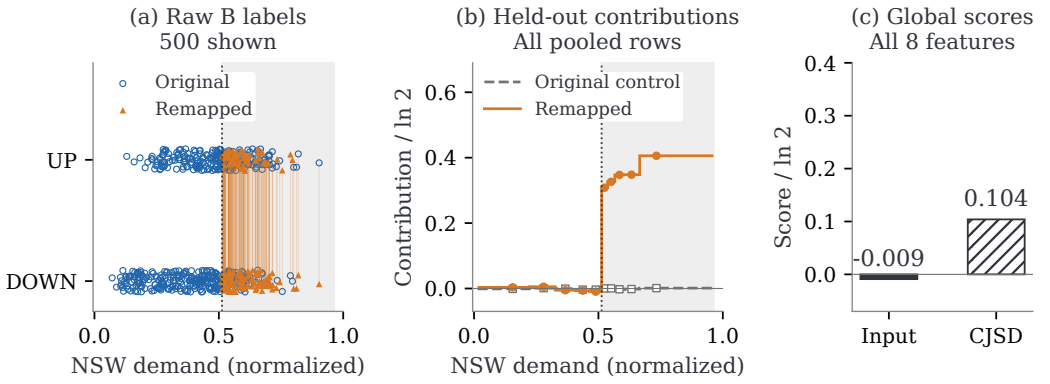

Figure 3: Electricity: records → regional contributions → global score. (a) A fixed 500-row subset of source B; circles are original labels and triangles their replacements, with small vertical jitter. (b) Mean paired held-out contributions in demand bands, using all 14,000 pooled rows and an original-label control on the same inputs and folds. Shading marks the imposed region. (c) Components use all eight input features and all rows. Count-weighted band means recover the global conditional score.

to $\widehat{D_{\mathrm{CJS}}}/\ln 2 \simeq 0.104$. Thus the score identifies where the fitted source prediction gains from labels. The relabeling fraction $2{,}096/7{,}000$ is not the same quantity: deterministic population labels are not established for these records, and the fitted score also contains learning error. Appendix B.3 contrasts this example with input-weighted sampling that retains the original labels.

4.3 What does a small score mean under strong input separation?

The gas-sensor drift benchmark contains 13,910 measurements of six gases, from 16 sensors across ten batches and 36 months (Vergara et al., 2012; Rodríguez-Luján et al., 2014; Vergara, 2012). Here X consists of 128 supplied response features, Y is recorded gas identity, and Z indicates the batch. Inputs and labels are unaltered. Concentration is used only for plotting. Histogram-based gradient boosting (HGB) classifiers (Friedman, 2001; scikit-learn developers, 2025) supply both components.

First versus last batch gives mean input score 0.965 and conditional score 0.004 (Figure 4). The corresponding mean held-out losses, divided by $\ln 2$, are 0.035 for T_1 and 0.031 for T_2, versus 1

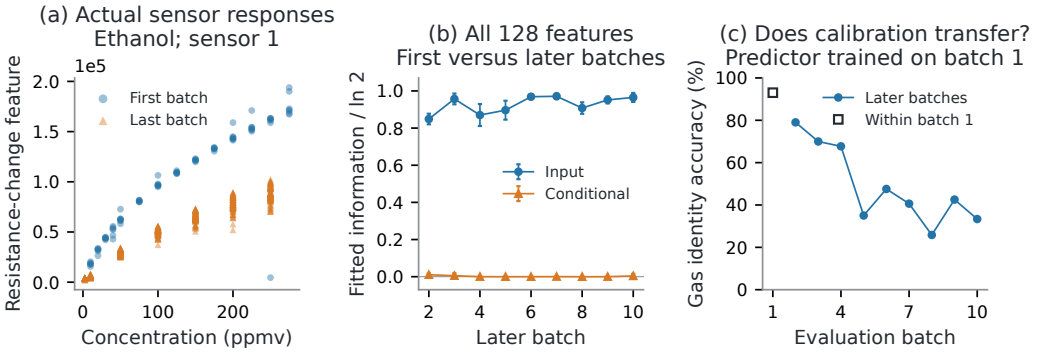

Figure 4: Unaltered gas measurements, their diagnosis, and transfer context. (a) Individual ethanol records: concentration and sensor 1's supplied resistance-change feature. (b) Full-array input and conditional scores, batch 1 versus each later batch; means and SDs over eight resampling/fold seeds. (c) Gas-identification accuracy: three-fold evaluation within batch 1 (open square), then fitting on all of batch 1 and evaluating later batches. Gas and concentration composition vary across batches. This observational comparison does not isolate a causal effect of sensor aging.

for a constant predictor. Meanwhile, gas-identification accuracy is 93.0% within the first batch and 33.4% on the last. The small conditional score plainly does not certify useful transfer. The strong source information in the inputs supplies the missing context: labels add little source information when responses already distinguish the periods. The population bound in Equation (4) explains this reading; fitted scores alone are not a certificate of population overlap or conditional equality.

The practical next step is to inspect coverage of the later operating conditions before interpreting the small conditional number. The true CJSD and a pure mechanism intervention are unknown here. All adjacent comparisons and within-batch controls, including unfavorable fitted control values, are in Appendix D. Within-batch fitted input scores reach -0.727 (input loss 1.727 in the same units): the HGB learner can overfit a random source label, which reinforces the need to inspect controls rather than read fitted values as population identities.

5 Same Quantity, Different Estimators

We compare direct source-classifier log loss with CCMI's MI-difference route (Mukherjee et al., 2019) and the mixed-variable nearest-neighbor estimator of Mesner & Shalizi (2021) (MS). Each estimates the same CJSD target. All receive the same 12,000 observations. The two neural routes match architecture, gradient-update count, and examples per update; neither is tuned against the known targets. HGB and MS share the observation budget rather than the neural compute budget.

The controlled laws have $d \in \{8, 128\}$, axis-aligned or shared rotated coordinates, and four conditions: unchanged, input shift, and deterministic label changes of population mass 0.03 or 0.30. Rotation changes the geometry seen by a learner while preserving the population target. Table 1 gives the high-dimensional rotated case, including both controls and the small signal. Appendix E reports every dimension, orientation, seed summary, and fixed k sensitivity.

Across all 128 condition/dimension/geometry/seed values per method, root mean squared error (RMSE) is 0.0391 for paired multilayer perceptrons (MLPs), 0.0689 for CCMI, 0.1071 for paired HGB, and 0.1134 for MS with $k = 5$. For the 0.30 change, paired MLP means span 0.231–0.267 across all four dimension/geometry settings. The same estimator misses the 0.03 change and has appreciable null variability. HGB preserves near-zero control means yet loses large-change magnitude under rotation and dimension. Near-zero null behavior is therefore a learner property, not a guarantee of the paired-loss route.

These results support the direct source objective as a useful way to estimate the diagnostic, with a learner suited to the input geometry. They do not imply uniform superiority of one estimator. The CCMI result uses a documented PyTorch port of its published MI-difference route, adapted to

Table 1: Same-target estimates on rotated $d = 128$ inputs. Mean $\pm$ seed SD over eight seeds, in bits; targets appear in the headings. Every method receives the same 12,000 observations. The neural routes also match architecture and update budget. Both nulls, small changes, and large changes are shown; SDs are not population CJSD confidence intervals.

Estimator	Null (0)	Input shift (0)	Rule (.03)	Rule (.30)
Paired HGB	-0.001 ± 0.003	-0.000 ± 0.002	0.000 ± 0.002	0.001 ± 0.004
Paired MLP	-0.010 ± 0.047	0.003 ± 0.029	-0.017 ± 0.026	0.231 ± 0.051
CCMI MI-diff.	-0.022 ± 0.051	0.000 ± 0.039	-0.015 ± 0.062	0.169 ± 0.063
MS ($k = 5$)	-0.003 ± 0.003	-0.004 ± 0.003	-0.001 ± 0.002	0.012 ± 0.004

common folds and mixed variables; the comparison covers this fixed-budget configuration rather than all CCMI variants.

6 Related work and scope

Dataset change. Concept-drift research already distinguishes changes in input and conditional distributions (Webb et al., 2016; Gama et al., 2014), including quantitative drift maps (Webb et al., 2018). We connect that diagnostic question to an additive information decomposition and its overlap weighting. Input two-sample methods such as maximum mean discrepancy (Gretton et al., 2012) and classifier two-sample tests (Lopez-Paz & Oquab, 2017) measure input differences; joint tests can signal a change without assigning it to the two components. Cross-learning scores instead compare fitted predictors' transfer (Sun et al., 2025).

Diagnosing performance changes. Distribution shift decomposition (DISDE) separates changes in a fixed predictor's performance due to input frequency, conditional outcomes, and limited support overlap (Cai et al., 2023). Distribution-based Shapley attribution likewise assigns performance changes to shifting distributions (Zhang et al., 2023). WhyShift supplies natural tabular shifts and identifies input regions with conditional change (Liu et al., 2023). Our population target describes the dataset pair independently of a task predictor or task loss; its estimate depends on source classifiers. It complements performance attribution without claiming exclusive regional diagnosis.

Conditional information and predictive losses. Conditional two-sample testing also uses the source-indicator reduction to $Z \perp Y \mid X$ (Lee et al., 2024; Hu & Lei, 2020). CCMI (Mukherjee et al., 2019) and mixed-data CMI estimators (Mesner & Shalizi, 2021) provide alternative estimation routes; none is a competing definition of CJSD. Predictive $\mathcal{V}$-information (Xu et al., 2020) and conditional probing (Hewitt et al., 2021) connect predictive log-loss differences to usable information. Equation (6) instead compares fixed fitted source classifiers with Shannon information, retaining both excess risks. Our local-MI neighborhood plug-in is another approximation to the same target, defined and delimited in Appendix B.2.

Scope of the reading. The diagnostic needs labeled datasets with common input and label meanings. Class-prior changes can alter $P(Y \mid X)$, so the conditional component does not identify a cause. Reading a small score requires adequate overlap and joint-classifier accuracy. Embeddings change the target to $I(Z; Y \mid \phi(X))$, which need not preserve raw-input conditional independence. Controlled changes validate known scenarios; gas supplies observational context without a known target. Record contributions guide inspection, not local significance or an optimal adaptation policy.

7 Conclusion

Two source-prediction questions separate input and conditional components, connect fitted label contributions to records, and expose the limits imposed by input overlap. Controlled records, natural sensors, and a common-target estimator comparison make this reading concrete. Reporting both components with learner sensitivity and controls turns a dataset discrepancy into a guide for inspection.

Reproducibility Statement

The appendix gives the population arguments, data constructions, and complete controls for the evidence retained in this paper. The anonymous supplement provides source code, per-seed results, figure data, verification scripts, source hashes, and instructions for rebuilding the manuscript. Its current README maps the focused claims to their evidence. Earlier experiments, including unsuccessful interventions, remain archived and are distinguished from the experiments supporting this manuscript. Recorded software versions and data retrieval instructions accompany the common-target and gas studies. Compact saved outputs permit numerical verification; omitted large synthetic prediction archives are identified by hashes and can be regenerated with the supplied commands.

Ethics Statement

The retained empirical evidence uses public electricity, forest-cover, and gas-sensor benchmarks and programmatically generated data. Conditional differences depend on the measured variables and shared support. The diagnostic supports examination of dataset changes; deployment decisions require domain evidence beyond these scores.

AI Use Statement

Large language model assistants were used to draft and check mathematical arguments, propose and critique experimental designs, implement and check analysis code, interpret results, identify literature, revise prose and figures, and organize the manuscript. AI-assisted simulated reviews informed the revision. Synthetic observations come from explicit programmatic generators, not LLM-generated data. No LLM was used for translation or qualitative data analysis. The authors are responsible for checking all claims, proofs, code, results, citations, and figures, and for the final submission.

References

G. I. Webb, R. Hyde, H. Cao, H.-L. Nguyen, F. Petitjean. Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30:964–994, 2016. https://doi.org/10.1007/s10618-015-0448-4.

G. I. Webb, L. K. Lee, B. Goethals, F. Petitjean. Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery*, 32:1179–1199, 2018. https://doi.org/10.1007/s10618-018-0554-1.

A. Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer, R. Huerta. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166–167:320–329, 2012. https://doi.org/10.1016/j.snb.2012.01.074.

I. Rodríguez-Luján, J. Fonollosa, A. Vergara, M. Homer, R. Huerta. On the calibration of sensor arrays for pattern recognition using the minimal number of experiments. *Chemometrics and Intelligent Laboratory Systems*, 130, 2014. https://doi.org/10.1016/j.chemolab.2013.10.012.

A. Vergara. Gas Sensor Array Drift at Different Concentrations. UCI Machine Learning Repository, 2012. https://doi.org/10.24432/C5MK6M.

S. Sun, H. H. Zhang, J. C. Watkins. Quantifying data similarity using cross learning. arXiv:2510.10866, 2025. https://arxiv.org/abs/2510.10866.

S. Mukherjee, H. Asnani, S. Kannan. CCMI: Classifier based conditional mutual information estimation. UAI 2019. https://proceedings.mlr.press/v115/mukherjee20a.html.

S. Lee, S. Cha, I. Kim. General frameworks for conditional two-sample testing. arXiv:2410.16636, 2024. https://arxiv.org/abs/2410.16636.

X. Hu, J. Lei. A two-sample conditional distribution test using conformal prediction and weighted rank sum. arXiv:2010.07147, 2020 (revised 2023). https://arxiv.org/abs/2010.07147.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola. A kernel two-sample test. *JMLR*, 2012. https://jmlr.org/papers/v13/gretton12a.html.

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*, 1991. https://doi.org/10.1109/18.61115.

D. Lopez-Paz, M. Oquab. Revisiting classifier two-sample tests. ICLR 2017. https://openreview.net/forum?id=SJkXfE5xx.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference*, 2000. https://doi.org/10.1016/S0378-3758(00)00115-4.

Y. Xu, S. Zhao, J. Song, R. Stewart, S. Ermon. A theory of usable information under computational constraints. ICLR 2020. https://arxiv.org/abs/2002.10689.

J. Hewitt, K. Ethayarajh, P. Liang, C. D. Manning. Conditional probing: measuring usable information beyond a baseline. EMNLP 2021, pp. 1626–1639. https://aclanthology.org/2021.emnlp-main.122/.

O. C. Mesner, C. R. Shalizi. Conditional mutual information estimation for mixed, discrete and continuous data. *IEEE Transactions on Information Theory*, 67(1):464–484, 2021. https://doi.org/10.1109/TIT.2020.3024886.

L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003. https://doi.org/10.1162/089976603321780272.

T. M. Cover, J. A. Thomas. *Elements of Information Theory*, 2nd ed. Wiley, 2006. https://doi.org/10.1002/047174882X.

J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, A. Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44, 2014. https://doi.org/10.1145/2523813.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. https://doi.org/10.1023/A:1010933404324.

J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001. https://doi.org/10.1214/aos/1013203451.

T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7.

F. Pedregosa et al. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830, 2011. https://jmlr.org/papers/v12/pedregosa11a.html.

scikit-learn developers. *scikit-learn 1.7 User Guide*, 2025. https://scikit-learn.org/1.7/user_guide.html.

M. Harries. Splice-2 comparative evaluation: Electricity pricing. Technical report UNSW-CSE-TR-9905, University of New South Wales, 1999. Dataset version used: Electricity, OpenML 151, version 1. https://www.openml.org/d/151.

J. Blackard. Covertype. UCI Machine Learning Repository, 1998. https://doi.org/10.24432/C50K5N.

J. Vanschoren, J. N. van Rijn, B. Bischl, L. Torgo. OpenML: Networked science in machine learning. arXiv:1407.7722, 2014. https://arxiv.org/abs/1407.7722.

O. Kallenberg. *Foundations of Modern Probability*, 3rd ed. Springer, 2021. https://doi.org/10.1007/978-3-030-61871-1.

D. P. Kingma, J. Ba. Adam: A method for stochastic optimization. ICLR 2015. https://arxiv.org/abs/1412.6980.

A. Paszke et al. PyTorch: An imperative style, high-performance deep learning library. NeurIPS 2019. https://arxiv.org/abs/1912.01703.

T. T. Cai, H. Namkoong, S. Yadlowsky. Diagnosing model performance under distribution shift. arXiv:2303.02011, 2023. https://arxiv.org/abs/2303.02011.

H. Zhang, H. Singh, M. Ghassemi, S. Joshi. "Why did the model fail?": Attributing model performance changes to distribution shifts. ICML, PMLR 202:41550–41578, 2023. https://proceedings.mlr.press/v202/zhang23ai.html.

J. Liu, T. Wang, P. Cui, H. Namkoong. On the need for a language describing distribution shifts: Illustrations on tabular datasets. NeurIPS Datasets and Benchmarks, 2023. https://doi.org/10.52202/075280-2237.

A PROOFS AND STATISTICAL DETAILS

Proof of Lemma 1. For a balanced mixture, $I(Z;X) = H(Z) - H(Z \mid X) = \ln 2 - \mathbb{E}h(w(X))$ is the usual Jensen–Shannon representation (Lin, 1991). Given $X = x$, the conditional law of Y is $m = w\eta_A + (1-w)\eta_B$. Therefore $I(Z;Y \mid X = x) = H(m) - wH(\eta_A) - (1-w)H(\eta_B) = \mathrm{JS}_w(\eta_A, \eta_B)$. Integration gives CJSD and the chain rule gives the joint decomposition. Nonnegativity follows from conditional mutual information, while $I(Z;Y \mid X) \le H(Z \mid X)$ gives the sharper upper bound. □

Proof of Proposition 1. For $0 < w < 1$, strict concavity of entropy implies $\mathrm{JS}_w(p,q) = 0$ if and only if $p = q$. At $w = 0$ or $w = 1$, the expression is zero for every pair. Since the integrand is nonnegative, its integral is zero exactly when it vanishes almost everywhere on the overlap. With mutually singular marginals $w \in \{0,1\}$ almost everywhere, so CJSD is always zero. No choice of a conditional version outside its task's support changes this conclusion. □

Proof of Proposition 2. Both conditional entropies are zero. If $f_A(x) = f_B(x)$, their mixture is a point mass with entropy zero. If the labels differ, the two mixture masses are $w(x)$ and $1 - w(x)$, giving entropy $h(w(x))$. Integrating proves the first formula. Equal marginals imply $w = 1/2$ almost everywhere, so $h(w) = \ln 2$, yielding the drift-mass law. □

Why the discrepancy is not a metric. Take $P_A = \delta_{(0,0)}$, $P_B = \delta_{(1,0)}$, and $P_C = \delta_{(0,1)}$ on binary (X,Y). Then $D_{\mathrm{CJS}}(A,B) = D_{\mathrm{CJS}}(B,C) = 0$ because the input supports are disjoint, whereas $D_{\mathrm{CJS}}(A,C) = \ln 2$. Thus even $\sqrt{D_{\mathrm{CJS}}}$ can violate the triangle inequality. The same example shows why equality of conditionals on overlap need not distinguish whole joint laws.

Fitted loss identity and its implications. Condition on a fitted probabilistic classifier T_j. Expanding its expected log loss as conditional entropy plus expected KL divergence gives $R_j = H(Z \mid U_j) + \epsilon_j$. Subtracting proves Equation (6). In particular,

$$D_{\mathrm{CJS}} - \epsilon_2 \le \widetilde{D} \le D_{\mathrm{CJS}} + \epsilon_1, \qquad |\widetilde{D} - D_{\mathrm{CJS}}| \le \max(\epsilon_1, \epsilon_2).$$

These are statements about fixed-model population risks. Evaluation noise is an additional term; the signs of individual empirical scores do not identify the signs or sizes of the two excess risks. One-sided certificates obtained by correcting the fitted conditional gap require justified bounds on those risks, including approximation error. Feature nesting alone supplies no such bound. The separate upper bound in Equation (7) follows directly from $\epsilon_1 \ge 0$ and $D_{\mathrm{CJS}} \le H(Z \mid X)$; it needs no approximation-error bound.

Cross-fitting and consistency. For a fixed number of source-stratified folds, let $a_k = N_k/N$, $\widetilde{D}_k$ be the fresh-data gap of fold k's fitted pair, and $\widehat{D}_k$ its held-out estimate. Then exactly

$$\widehat{D_{\mathrm{CJS}}} - D_{\mathrm{CJS}} = \sum_k a_k(\widehat{D}_k - \widetilde{D}_k) + \sum_k a_k(\epsilon_{1,k} - \epsilon_{2,k}).$$

Under independent observations and fitting confined to the other folds, foldwise evaluation errors vanish as fold sizes grow if the clipped loss bounds grow slowly enough. If both excess risks also vanish in probability, the estimate is consistent. A fixed clip is compatible with this condition when the true posteriors stay inside that clip; otherwise a vanishing clip ϵ_N may be used with $\ln^2(1/\epsilon_N)/\min_k N_k \to 0$. These are conditional guarantees, not properties demonstrated for each learner in our finite experiments. Shared training observations make fold estimates dependent; their naive paired standard error is not a population CJSD confidence interval.

B CONTROLLED RECORDS AND THEIR PROTOCOLS

B.1 THE 202-PAIR BATTERY

The entry point is `exp15_compare.py`, with data loaders in `cpdlib/data_real.py`. Electricity uses OpenML version 1 (Vanschoren et al., 2014) and Covertype uses scikit-learn's public dataset loader (Pedregosa et al., 2011). Electricity categorical columns use integer codes. No learned embedding is used. Both classifiers are 200-tree random forests, minimum leaf size 20, with three shared source-stratified folds over 14,000 pooled rows (approximately 9,333 fitting rows per fold).

The synthetic battery has 13 conditions and ten seeds per condition: null; input mean shifts $s \in \{1, 2, 3\}$; label-rule rotations $\theta \in \{10, 20, 45\}$ degrees; both $(s, \theta) = (2, 45)$; and local probability flips of requested mass $\rho \in \{.05, .1, .2, .3, .5\}$. For $X \sim N((s, 0), I_2)$, set

$$u_1 = X_1 \cos\theta - X_2 \sin\theta, \quad u_2 = X_1 \sin\theta + X_2 \cos\theta, \quad P(Y = 1 \mid X) = \sigma(10[u_2 - \sin u_1]),$$

where $\sigma(t) = 1/(1 + e^{-t})$. Local changes replace this probability by its complement where $X_1 > \Phi^{-1}(1 - \rho)$, with Φ the standard normal cumulative distribution function. A label-rule rotation changes the conditional law; it is different from rotating both datasets' coordinates after generating labels in Appendix E.

Each real dataset has six conditions and six seeds: null, input shift, regional relabeling at requested masses .1, .3, .5, and both at .3. Null and label-change pairs use disjoint uniform samples. Input-shift pairs use opposite exponential sampling weights on an approximate first principal component of standardized features (Hastie et al., 2009); the second sample excludes rows used in the first. Label change cyclically remaps classes above an empirical feature quantile. Its feature is selected among columns with more than ten unique values in the construction subset. Ties can make realized mass differ from the requested mass. Each source contains 7,000 rows, pooling the original 4,000/3,000 benchmark split for cross-fitting. Seeds reuse finite real-record pools rather than independently collected environments.

The reported ranking compares 80 synthetic and 18 per-real-family label-change pairs against 30 synthetic and six per-real-family covariate-only pairs. Null and combined conditions remain in the plotted 202 pairs but are excluded from that ranking. The archived benchmark contains additional input-distance and task-predictor scores; they are not used as estimators of the same information target in the main comparison.

B.2 LOCAL-MI: AN ALTERNATIVE PLUG-IN, NOT A COMPETING TARGET

Our local-MI implementation replaces probabilities by empirical frequencies in the discrete mutual-information formula (Paninski, 2003). On at most 4,000 pooled rows, standardize coordinates by their pooled mean and standard deviation plus 10^{-9}. For each input x_i, take the $m = k + 1$ nearest input rows in Euclidean distance, with $k = 15$. The row itself is included for untied distances; ties follow the library. Let $\widehat{p}_i(z, y)$ be the source–label frequency table in that neighborhood. The estimate in nats is

$$\widehat{I}_{\text{local}} = \frac{1}{M} \sum_{i=1}^{M} \sum_{z,y} \widehat{p}_i(z, y) \ln \frac{\widehat{p}_i(z, y)}{\widehat{p}_i(z)\widehat{p}_i(y)},$$

with zero-frequency terms omitted. Neighborhoods approximate conditioning on inputs. This author-implemented plug-in also attains AUC 1.00 in each battery family, but finite neighborhood tables can have positive MI under independence. No bias correction or consistency claim is made for fixed k. It is distinct from the published mixed-variable estimator of Mesner & Shalizi (2021); the paired-classifier excess-risk identity does not apply to its estimation error. Code and complete archived results are in `exp26_new_baselines.py` and `results/exp26/`.

B.3 ELECTRICITY: INPUT SAMPLING VERSUS CHANGED LABELS

Figure 5 retains a full visual comparison for seed 0. Each source has 7,000 disjoint records. Input weighting retains original UP/DOWN labels. The other pair uses disjoint uniform samples and exchanges UP/DOWN in B where `nswdemand`> 0.512943, changing 2,096 labels. The threshold

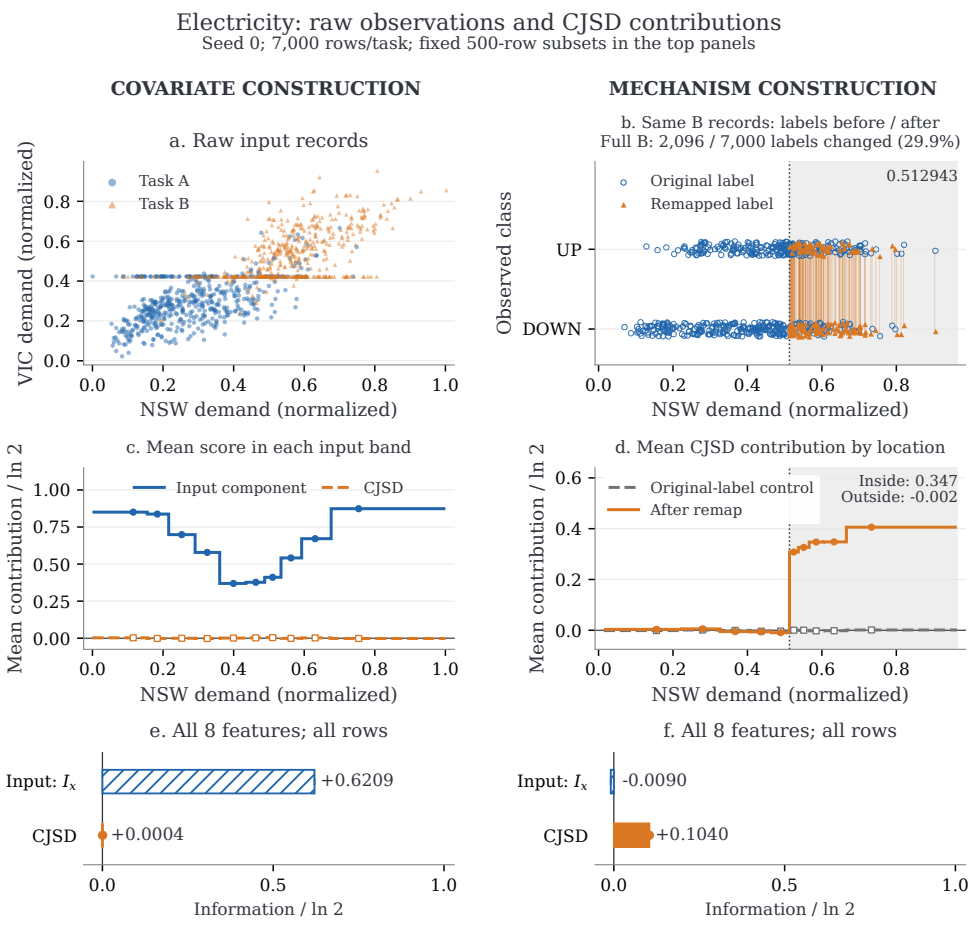

Figure 5: Two controlled comparisons of actual Electricity records. (a) Raw demand coordinates for 500 fixed-position rows per source. (b) The same 500 B records before/after label remapping, with connecting segments and label jitter. (c,d) All-row held-out contributions averaged within ten input bands. (e,f) Global RF scores using all eight features. Shading marks the imposed remapping region; columns are separate constructions, not different historical periods.

is a construction choice, not a discovered boundary. The source-classifier folds and original-label control use identical inputs; the control refits on the original labels.

For each row define $a_i = (\ln 2 - \ell_i^{(1)})/\ln 2$ and $c_i = (\ell_i^{(1)} - \ell_i^{(2)})/\ln 2$. Count-weighted band means recover the corresponding global score exactly. Input weighting gives $(\widehat{I}_x, \widehat{D_{\mathrm{CJS}}})/\ln 2 = (0.6209, 0.0004)$. After remapping, the conditional mean is 0.3471021 inside the region and -0.0024729 outside. Of 14,000 pooled rows, 4,263 are inside (mass 0.3045), giving 0.1039727 globally (input score -0.0090). These are contributions of the fitted model, not population local CJSDs. The original-label control stays near zero in both regions. Saved per-row losses, selected rows, and thresholds support figure regeneration without refitting in `paper/figure_data/`.

C FROM SYNTHETIC LAWS AND SAMPLES TO THE TWO COMPONENTS

The introductory illustration. Figure 1 uses $X_A \sim N(0,1)$ and $P_A(Y=1 \mid x) = \sigma(3x)$. Dataset B either shifts the input mean to 1.5, changes the conditional to $\sigma(-3x)$, or does both. Numerical integration gives normalized components approximately $(0.319, 0)$, $(0, 0.480)$, and $(0.319, 0.370)$. The conditional score decreases in the combined case because source uncertainty at a given input decreases. These are explanatory laws, separate from the benchmark. The builder `make_main_figures.py` checks the integrals using independent quadrature routines.

C.1 PROBABILITY LAWS IN THE BATTERY

The reference in Figure 6 has $s = \theta = 0$ under Appendix B's nonlinear Bernoulli law. Five comparisons change neither law, only s to 2, only θ to $45°$, both, or complement the label probability above $x_1 > \Phi^{-1}(.7)$. The local region has population mass 0.3; because labels are stochastic, that mass is not its CJSD.

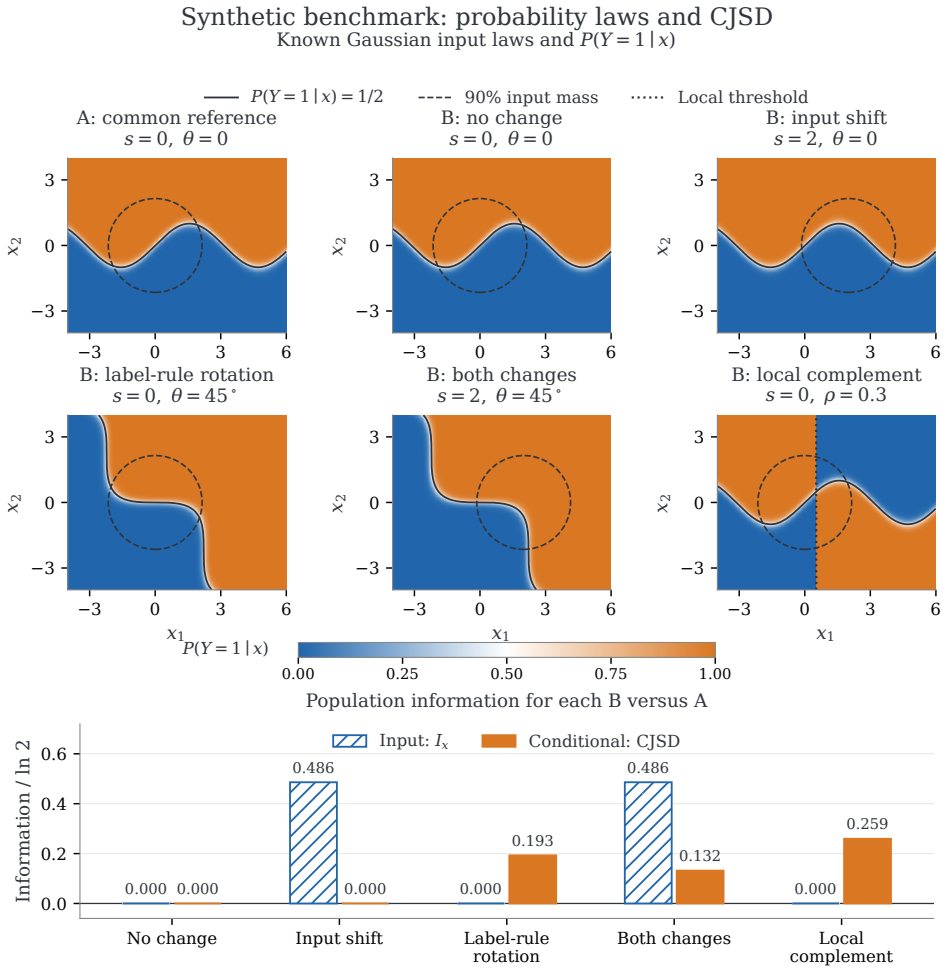

Figure 6: Known laws and their population components. Maps share a probability scale; dashed circles enclose 90% of Gaussian input mass, solid curves mark label probability $1/2$, and the dotted line marks the local change. All input laws have full support. Bottom bars compare B with the common A. Integration with 384 and 640 quadrature nodes agrees within 10^{-9}. Input shift alone gives CJSD zero; the three label-rule changes give 0.193, 0.132, and 0.259.

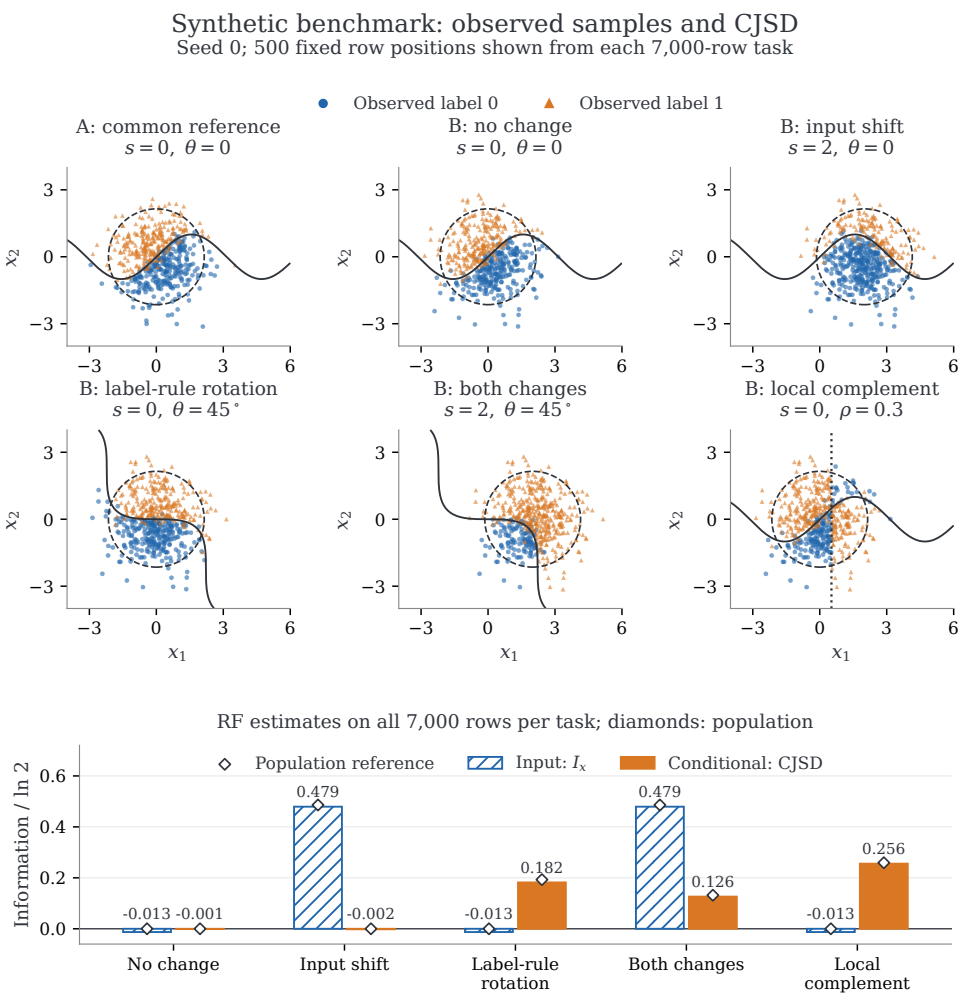

Figure 7: The same five benchmark conditions, as generated observations followed by fitted results. Fixed displayed subsets make the point clouds readable; classifiers use all 7,000 observations per source. The lower panels contrast seed-0 RF estimates with the population references. Differences reflect finite fitting and evaluation rather than a change of target. Source identities and saved scores are recorded in `synthetic_shift_examples.json`.

C.2 Generated observations and fitted scores

The visual sequence separates three objects: the known generating law, the sampled records, and the fitted diagnostic. Probability maps reveal where the rule changed; input clouds show which regions were sampled; the bars show what the two classifiers recovered. All five population references also appear as black crosses in Figure 2.

D OBSERVED GAS-SENSOR CHANGE: COMPLETE CONTROLS

D.1 MEASUREMENTS, LABELS, AND THE QUESTION BEING ASKED

The UCI Gas Sensor Array Drift Dataset at Different Concentrations contains 13,910 measurements from 16 chemical sensors, with eight supplied response features per sensor, six recorded gas identities, and ten chronological batches spanning 36 months (Vergara et al., 2012; Rodríguez-Luján et al., 2014; Vergara, 2012). The original study documents the practical problem of calibration transfer over sensor drift. Here X comprises the 128 supplied response features, Y is the recorded gas identity, and Z indicates the compared batch. Concentration is used for descriptive plots and is excluded from X. No recorded input or gas label is altered. These features summarize each measurement's steady and transient responses; they are not raw time-series waveforms. Feature 1 is the first sensor's resistance-change response.

Figure 8 displays individual records before aggregation. For example, the ethanol response differs between the first and last batches even at concentrations represented in both. This makes the measurement change concrete, but is not by itself evidence about $P(Y \mid X)$: the gas and concentration mix also varies over time. Our statistical comparison asks what the recorded gas identity adds to identifying the period after seeing the sensor responses. It does not attribute that difference causally to aging, nor assume a known population CJSD for this dataset.

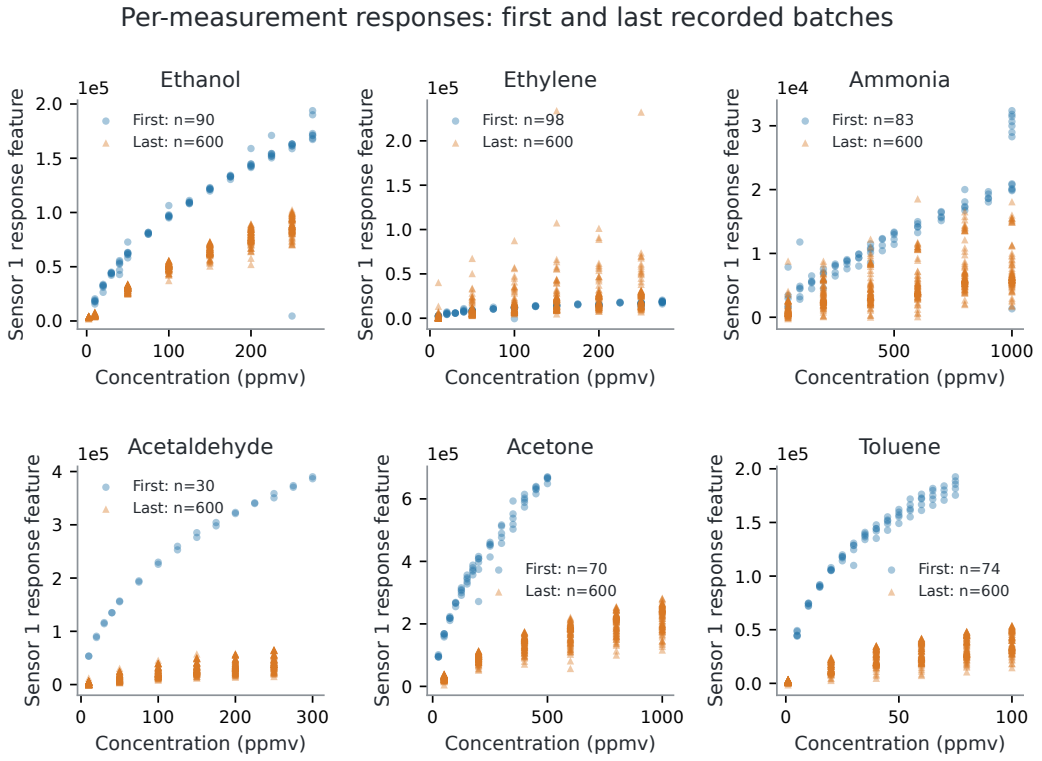

Figure 8: Actual gas-sensor records, with no relabeling or injected shift. Every point is one supplied response feature and its recorded concentration; facets retain the six original gas identities. The first and last batches are fixed chronological endpoints. Counts show all available records in each facet, before the source balancing used for estimation.

D.2 COMPARISONS, CONTROLS, AND PRACTICAL CONTEXT

We compare all nine adjacent batch pairs, batch 1 with every later batch, and random disjoint source splits within each of the ten batches. Counting the first–second comparison once gives 27 settings. We use eight resampling/fold seeds per setting, for 216 fits of the paired procedure. For different

batches, the smaller batch sets the sample size per source; we sample equally many rows from the larger batch without replacement. For a within-batch control, a random permutation is split into two disjoint halves. We retain observed class composition; class balancing would change the dataset law being compared. An audit found no nonfinite features or duplicate full response vectors among the 13,910 records.

Both source classifiers are histogram-based gradient boosting models (HGB, 150 iterations, no early stopping). The joint classifier receives six one-hot gas indicators in addition to X. Three source-stratified folds supply held-out probabilities; identical observed feature vectors, if present in a sensor view, are grouped within a fold. Probabilities are clipped to $[10^{-4}, 1 - 10^{-4}]$ before log loss. We retain signed scores. Table 2 reports every setting, including negative fitted input scores on within-batch controls: flexible source classifiers can predict a random source label worse than the constant predictor. These control scores expose fitting error rather than negative population information. Whiskers are SDs over resampling/fold seeds of the same recorded dataset, not independent temporal replications or CJSD confidence intervals. Gas/record dependence is not modeled as an inferential process.

The first–last mean scores are 0.9650 for the input component and 0.0041 for the conditional component, normalized by $\ln 2$. All nine adjacent comparisons also have substantial input separation. The small additional source information in the labels should therefore be read together with the strong source information in the responses. At the population level $D_{\mathrm{CJS}} \leq H(Z \mid X)$; a fitted near-maximal input score motivates checking overlap and classifier error, rather than substituting the fitted residual for a proved population bound. The old gas predictor's accuracy is 93.0% within the first batch and 33.4% on the last batch. Thus a small conditional score is compatible with poor transfer when input coverage has changed. The two-component report makes this interpretation visible, even without ground-truth conditional discrepancy or an isolated physical intervention.

Table 2: All full-array gas comparisons and within-batch random-source controls. Mean $\pm$ SD across eight resampling/fold seeds; normalized by $\ln 2$. These are fitted scores, so negative values are retained.

Batches	n/source	$\widehat{I}_x$	$\widehat{D}$
1-1	222	-0.7267 ± 0.1085	-0.0017 ± 0.0068
1-2	445	0.8488 ± 0.0287	0.0104 ± 0.0071
1-3	445	0.9569 ± 0.0291	0.0054 ± 0.0131
1-4	161	0.8702 ± 0.0591	0.0001 ± 0.0021
1-5	197	0.8964 ± 0.0509	-0.0000 ± 0.0018
1-6	445	0.9688 ± 0.0105	-0.0000 ± 0.0001
1-7	445	0.9713 ± 0.0149	0.0004 ± 0.0020
1-8	294	0.9079 ± 0.0309	-0.0002 ± 0.0005
1-9	445	0.9517 ± 0.0190	0.0001 ± 0.0002
1-10	445	0.9650 ± 0.0242	0.0041 ± 0.0117
2-2	622	-0.5983 ± 0.0415	0.0008 ± 0.0047
2-3	1244	0.9672 ± 0.0061	-0.0019 ± 0.0034
3-3	793	-0.3833 ± 0.0240	-0.0012 ± 0.0029
3-4	161	0.8262 ± 0.0310	-0.0003 ± 0.0066
4-4	80	-0.4941 ± 0.1914	0.0000 ± 0.0000
4-5	161	0.8034 ± 0.0488	0.0007 ± 0.0040
5-5	98	-0.5183 ± 0.0965	0.0000 ± 0.0000
5-6	197	0.8641 ± 0.0612	-0.0061 ± 0.0118
6-6	1150	-0.2914 ± 0.0190	0.0024 ± 0.0034
6-7	2300	0.9698 ± 0.0071	-0.0008 ± 0.0014
7-7	1806	-0.1831 ± 0.0224	-0.0014 ± 0.0043
7-8	294	0.8662 ± 0.0278	-0.0024 ± 0.0091
8-8	147	-0.6515 ± 0.1530	0.0010 ± 0.0028
8-9	294	0.9285 ± 0.0245	0.0003 ± 0.0019
9-9	235	-0.6217 ± 0.1516	0.0000 ± 0.0000
9-10	470	0.9679 ± 0.0243	0.0005 ± 0.0011
10-10	1800	-0.1971 ± 0.0195	0.0021 ± 0.0035

E SAME-TARGET COMPARISON WITH ESTABLISHED CMI ESTIMATORS

E.1 TARGET, OBSERVATIONS, AND BUDGET

This experiment compares estimators of the same Shannon quantity $I(Z;Y \mid X)/\ln 2$, rather than comparing a discrepancy magnitude with a conditional-test p-value. We include the paired log-loss route with HGB and an MLP, the classifier-based CCMI MI-difference route (Mukherjee et al., 2019), and the mixed-variable nearest-neighbor estimator of Mesner & Shalizi (2021) (MS). The latter is distinct from our local-MI neighborhood plug-in; both are estimators of the CJSD target, not alternative definitions of dataset discrepancy.

The known-target laws use $X_A \sim N(0, I_d)$ and $Y_A = \mathbf{1}\{X_1 > 0\}$. The null keeps this law for B. Input shift sets $X_B \sim N(1.5u, I_d)$ for a seeded unit vector u while retaining the label rule. Rule change uses equal input laws and flips B's label where $X_2 > \Phi^{-1}(1-\rho)$, for $\rho = .03$ or $.30$. Common orthogonal rotations are applied to both input arrays after label generation. These invertible transformations preserve the target. We use 6,000 observations per source (12,000 total), $d \in \{8, 128\}$, axis-aligned and shared orthogonally rotated inputs, and seeds 0–7. There are four conditions: unchanged law (target 0), input shift with unchanged rule (0), and equal-input deterministic rule changes with population disagreement masses 0.03 and 0.30. All 32 geometry/dimension/seed cells and all four conditions are reported. The shared source-stratified three folds contain 8,000 fitting and 4,000 held-out observations each. MS uses all 12,000 observations, as appropriate for its sample-based estimator; it has no fitting/evaluation split. No method is tuned against known target values or evaluated by choosing a favorable seed.

E.2 IMPLEMENTATIONS AND ADAPTATIONS

Paired log loss. HGB uses 150 iterations without early stopping. The MLP has two 64-unit rectified-linear (ReLU) layers (Hastie et al., 2009). Weights are initialized uniformly on $[-\sqrt{6/(f_{\text{in}} + f_{\text{out}})}, \sqrt{6/(f_{\text{in}} + f_{\text{out}})}]$, where $f_{\text{in}}, f_{\text{out}}$ are layer widths, with zero biases. Adam (Kingma & Ba, 2015) uses learning rate 0.001, and a weights-only penalty of $0.001\|W\|_2^2/2$. Both use exactly the same outer folds and retain signed loss differences. The input-only model has two zero padding coordinates where joint models receive one-hot binary labels. Neural preprocessing uses only fitting rows to standardize features; zero-variance coordinates have scale one. Within a fold, the six models needed across all conditions have separate parameters and share an initialization for paired comparison. Vectorizing these independent models on the GPU does not share learned weights.

CCMI MI-difference. For $V = X$ and $V = (X, Y)$, we train a classifier to discriminate the joint law of (Z, V) from its product surrogate, then evaluate the Donsker–Varadhan (DV) score

$$\widehat{I}_{\text{DV}}(Z;V) = \overline{\log r}_{\text{joint}} - \log \overline{r}_{\text{product}}, \qquad r = \frac{\widehat{p}}{1-\widehat{p}}.$$

The conditional estimate subtracts the two MI estimates. We implemented this published route in PyTorch (Paszke et al., 2019) after inspecting the authors' TensorFlow reference code. This is a documented port, not an execution of the original binary. The same MLP architecture, optimizer, regularizer, probability clipping, and fixed update count are used for both neural routes. Each outer-fold model receives $\lfloor 20 n_{\text{train}}/64 \rfloor$ updates: CCMI uses 64 joint and 64 product rows per update, and the paired log-loss route uses 128 source-labeled observations. Thus available observations, architecture, examples per update, gradient-update count, and zero tuning budget match; different objectives need not take identical wall time. HGB shares the observation budget, not the neural compute budget.

Product surrogates permute Z within the fitting or held-out partition separately, preserving the marginal distributions without moving records across folds. We adapt the reference pipeline to shared cross-fitting, fitting-only standardization, binary-source/categorical-label encoding, and signed outputs. Log odds are capped by the same probability clipping as the paired route. The product-law DV objective applies to these mixed variables; we do not claim that this adaptation inherits every assumption or consistency theorem of the original continuous-data analysis. No neural early stopping or evaluation-based hyperparameter selection is used.

Mixed-variable nearest neighbors. Our memory-efficient distance implementation follows the authors' inclusive-radius counting convention, using Chebyshev (maximum-coordinate) distance, distance one for unequal categories, and marginal empirical ranks in $(0, 1)$ for continuous coordinates. With neighbor radius ρ_i in the joint (Z, Y, X) space, it computes

$$\frac{1}{n}\sum_i \{\psi(k_i) - \psi(n_{ZX,i}) - \psi(n_{YX,i}) + \psi(n_{X,i})\}/\ln 2,$$

where all counts exclude the point itself and include boundary ties; k_i is the joint count and ψ is the digamma function. We fix $k = 5$ as primary and report $k = 10, 20$ as prespecified sensitivities. These numerical-distance choices do not make the estimator invariant to arbitrary rotations. We validated the implementation against the authors' code on small mixed-variable datasets with and without continuous ties; maximum absolute discrepancy was below 10^{-14}. A separate known-law check verifies the classifier DV arithmetic, and saved held-out logits allow independent recomputation of every neural score.

E.3 Results: recovery, controls, and remaining attenuation

The paired MLP reaches large-change means 0.2525, 0.2502, 0.2673, and 0.2310 across $(8, \text{axis})$, $(8, \text{rot})$, $(128, \text{axis})$, and $(128, \text{rot})$, respectively, against true 0.30. Corresponding CCMI means range from 0.1688 to 0.2028; MS with $k = 5$ declines from 0.1859 in the low-dimensional axis-aligned case to about 0.01 at $d = 128$. HGB is closest in the $d = 8$ axis-aligned case (0.2611), but attenuates strongly with dimension and rotation. Thus the paired objective can retain substantial rotated signal when paired with a suitable learner, without claiming that the tree route itself is rotation invariant.

Equal-weight RMSE over all 128 condition/dimension/geometry/seed scores is 0.0391 for paired MLP, 0.0689 for CCMI, 0.1071 for paired HGB, and 0.1134 for MS ($k = 5$). The main table includes all four conditions for rotated $d = 128$; these tables retain every condition and geometry. Paired MLP underestimates the 0.03 change and has negative high-dimensional small-change means. At $d = 128$, its null means are -0.0131 and -0.0098, with SDs 0.0287 and 0.0473. CCMI also has appreciable null variation. Near-zero tree control means therefore do not establish a property of the information-estimation route for every learner. These are fixed-budget comparisons, not tuned best-case performance bounds for any method.

Reproducibility. The supplement includes the prespecified protocol, source commits and archive hashes, the PyTorch port, author-code validation, per-seed scores, input/fold hashes, timings, software versions, and table/figure builder. Prediction archives are retained with SHA-256 checksums and regenerated by the supplied commands. Raw UCI measurements and upstream implementations are retrieved from their cited sources rather than copied into the archive. Gas results are observations on one longitudinal dataset; the 32 synthetic cells and sensor views are not additional independently collected datasets.

Table 3: Same-target estimates: $d = 8$, axis. Mean $\pm$ seed SD; eight seeds; normalized by $\ln 2$.

Estimator	Null	Input shift	Rule .03	Rule .30
CJSD-HGB	0.0001 ± 0.0004	0.0002 ± 0.0004	0.0165 ± 0.0029	0.2611 ± 0.0107
CJSD-MLP	-0.0012 ± 0.0008	-0.0014 ± 0.0013	0.0039 ± 0.0015	0.2525 ± 0.0073
CCMI-MI-diff	-0.0000 ± 0.0002	-0.0012 ± 0.0013	-0.0001 ± 0.0002	0.2028 ± 0.0077
MS-k5	-0.0016 ± 0.0016	-0.0034 ± 0.0023	0.0057 ± 0.0028	0.1859 ± 0.0053
MS-k10	-0.0015 ± 0.0010	-0.0027 ± 0.0019	0.0040 ± 0.0020	0.1733 ± 0.0039
MS-k20	-0.0016 ± 0.0006	-0.0023 ± 0.0025	0.0022 ± 0.0014	0.1582 ± 0.0039

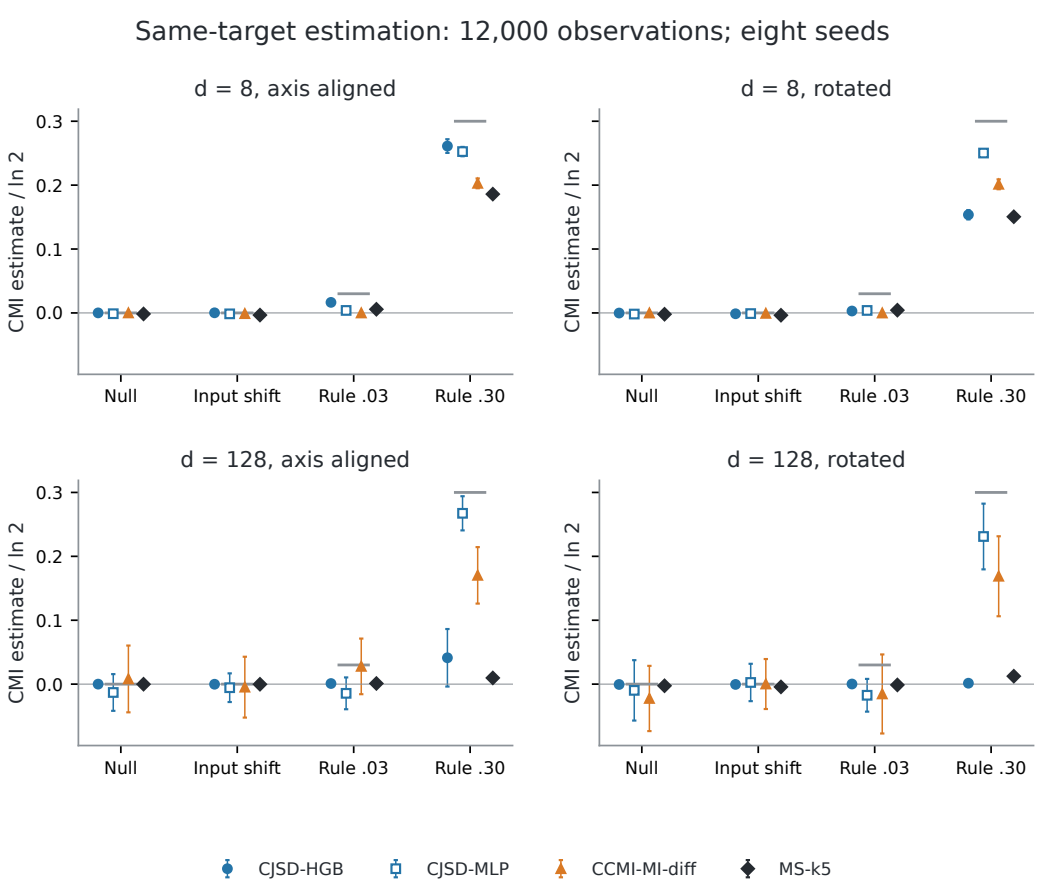

Figure 9: Common-target magnitude comparison. Each panel reports all four conditions for one dimension and geometry. Points and whiskers are means and SDs over eight generator seeds; gray ticks mark the population targets. All methods receive 12,000 observations. The two neural routes match architecture and gradient-update budget. The error analysis concerns signed estimates, so negative values remain visible.

Table 4: Same-target estimates: $d = 8$, rot. Mean $\pm$ seed SD; eight seeds; normalized by $\ln 2$.

Estimator	Null	Input shift	Rule .03	Rule .30
CJSD-HGB	-0.0002 ± 0.0014	-0.0013 ± 0.0022	0.0030 ± 0.0025	0.1537 ± 0.0073
CJSD-MLP	-0.0018 ± 0.0011	-0.0010 ± 0.0010	0.0039 ± 0.0017	0.2502 ± 0.0053
CCMI-MI-diff	-0.0002 ± 0.0005	-0.0007 ± 0.0011	-0.0002 ± 0.0005	0.2013 ± 0.0078
MS-k5	-0.0021 ± 0.0011	-0.0036 ± 0.0015	0.0043 ± 0.0013	0.1505 ± 0.0055
MS-k10	-0.0014 ± 0.0008	-0.0026 ± 0.0017	0.0040 ± 0.0013	0.1408 ± 0.0060
MS-k20	-0.0008 ± 0.0010	-0.0028 ± 0.0027	0.0033 ± 0.0014	0.1290 ± 0.0063

Table 5: Same-target estimates: $d = 128$, axis. Mean $\pm$ seed SD; eight seeds; normalized by $\ln 2$.

Estimator	Null	Input shift	Rule .03	Rule .30
CJSD-HGB	0.0000 ± 0.0000	-0.0002 ± 0.0005	0.0010 ± 0.0031	0.0412 ± 0.0450
CJSD-MLP	-0.0131 ± 0.0287	-0.0056 ± 0.0224	-0.0144 ± 0.0249	0.2673 ± 0.0267
CCMI-MI-diff	0.0081 ± 0.0523	-0.0047 ± 0.0476	0.0278 ± 0.0435	0.1702 ± 0.0442
MS-k5	-0.0001 ± 0.0034	-0.0002 ± 0.0043	0.0010 ± 0.0028	0.0097 ± 0.0032
MS-k10	0.0003 ± 0.0033	0.0007 ± 0.0020	0.0007 ± 0.0032	0.0082 ± 0.0025
MS-k20	0.0005 ± 0.0024	0.0005 ± 0.0014	0.0010 ± 0.0026	0.0067 ± 0.0020

Table 6: Same-target estimates: $d = 128$, rot. Mean $\pm$ seed SD; eight seeds; normalized by $\ln 2$.

Estimator	Null	Input shift	Rule .03	Rule .30
CJSD-HGB	-0.0005 ± 0.0035	-0.0005 ± 0.0023	0.0002 ± 0.0019	0.0014 ± 0.0043
CJSD-MLP	-0.0098 ± 0.0473	0.0026 ± 0.0292	-0.0174 ± 0.0256	0.2310 ± 0.0514
CCMI-MI-diff	-0.0224 ± 0.0510	0.0002 ± 0.0391	-0.0154 ± 0.0618	0.1688 ± 0.0626
MS-k5	-0.0026 ± 0.0026	-0.0043 ± 0.0026	-0.0012 ± 0.0022	0.0124 ± 0.0044
MS-k10	-0.0021 ± 0.0015	-0.0029 ± 0.0015	-0.0010 ± 0.0017	0.0118 ± 0.0030
MS-k20	-0.0017 ± 0.0020	-0.0020 ± 0.0018	-0.0010 ± 0.0022	0.0100 ± 0.0025

Table 7: Errors equally weight all four conditions, dimensions, geometries, and eight seeds (128 scores per estimator). The last column gives the largest absolute cell mean among the null and covariate controls. These weights are a reporting convention, not a model-selection objective. A constant-zero estimator has RMSE 0.1507 under this convention.

Estimator	Signed bias	RMSE	Max. absolute control mean
CJSD-HGB	-0.0528	0.1071	0.0013
CJSD-MLP	-0.0234	0.0391	0.0131
CCMI-MI-diff	-0.0366	0.0689	0.0224
MS-k5	-0.0606	0.1134	0.0043

Table 8: Observed mean wall seconds per complete four-condition seed run, including three outer folds for classifier routes. Neural models are vectorized on one RTX A6000; HGB uses two CPU threads. MS time includes all three neighbor counts and shares distance calculations. Different hardware kernels, sharing, and background CPU work make these execution records, not hardware-independent speed rankings.

Estimator	8 axis	8 rot.	128 axis	128 rot.
CJSD-HGB	6.2	6.2	22.4	22.3
CJSD-MLP	18.4	17.4	18.5	18.1
CCMI-MI-diff	18.8	17.8	19.0	18.8
MS, all three k	2.3	2.3	2.8	2.8