

Supplementary methods and validation

Comparing mutual information across unequal samples

Nick Barua and Koji Oda

Corresponding author: Nick Barua · s.nick.barua@gmail.com

Corresponding author affiliation 1: Department of Legal Medicine, Shiga University of Medical Science, Setatsukinowacho, Otsu 520-2192, Shiga, Japan

Corresponding author affiliation 2: AN Holdings Co., 2-18 Wakamatsu-cho, Nishinomiya, Hyogo 662-0035, Japan

Journal: Journal of Statistical Theory and Practice

Online Resource 1

This supplement presents the mathematical identities, implementation checks and original study record. It also retains the earlier simulation and stellar-catalogue analyses and describes the subsequent two-anchor moderate-imbalance extension in Section S7. We distinguish the original development and reserved stages from the retrospective audit and independently seeded extension. Online Resource 2, available at <https://doi.org/10.5281/zenodo.23466586>, contains the full cell summaries, procedure-level replicate outputs, code, configurations and provenance. Individual estimator-call checkpoints are omitted from the compact archive and can be regenerated from the supplied code and seed specifications. Earlier outcomes were known when the extension was specified; local freezes are not public preregistration. Archive paths are relative to the extracted root of Online Resource 2 unless a containing directory or nested ZIP is specified. Folder names are retained for reproducibility.

S1 Mathematical identities

S1.1 The finite sample permutation reference

Let a sample of N observations be drawn independently from a joint distribution P , and let π be an independent uniform permutation of the D rows. Suppose the statistic T is integrable. Define $a_N(P)$ as its sampling expectation and $c_N(P)$ as the sampling expectation of its conditional permutation mean. Averaging any positive number B of independent permutations has that same conditional expectation. Therefore

$$E_P[T_{N,B}^c] - I(P) = a_N(P) - c_N(P) - I(P).$$

Under independence, P equals the product of its two block marginals. Permuting one block then preserves the joint sampling distribution, including any dependence among coordinates within either block. Consequently $a_N(P)$ and $c_N(P)$ are equal, and the centred statistic has mean zero. Under an alternative, $c_N(P)$ remains an empirical-sample permutation reference; it need not equal the mean statistic on a fresh iid sample from the product population. Null centring therefore supplies no general alternative-unbiasedness result.

Writing b^c for the mean centred error, linearity gives the comparison identity

$$E[(T_B^c - T_A^c) - (I_B - I_A)] = b^c(N_B, P_B) - b^c(N_A, P_A).$$

Equal information alone does not make the right-hand side zero. This identity holds even if samples A and B are dependent. We use independent A and B samples within each simulated pair; the paired Monte Carlo intervals require independence between replicate pairs. These expectation identities distinguish null centring from estimation bias.

S1.2 Common size subsampling

Choose n_0 distinct indices uniformly from an iid parent sample of size N , independently of the observations. Unconditionally, the selected observations have the distribution of n_0 iid draws from the same population.

Applying the same preprocessing, statistic and permutation scheme consequently gives the ordinary n_0 -sample expectation. Averaging M random subsets does not alter it. Thus two identical joint distributions have equal expected common-size estimates; equality of their information values alone is insufficient.

Let Y_1 be a subset estimate, with independently drawn subset indices and permutation streams conditional on its parent sample S . Define $m(S)$ as its conditional mean. The variance decomposition is

$$\text{Var}\left[\frac{1}{M} \sum_{l=1}^M Y_l\right] = \text{Var}[m(S)] + \frac{1}{M} E[\text{Var}(Y_l | S)].$$

Increasing M reduces only the second term. Overlapping subsets do not create independent parent samples, and common-size averaging need not improve the mean squared error of a comparison. Nested prefixes also share their parent; resampling must preserve this dependence.

S1.3 The Gaussian covariance correction

For N iid nonsingular m -variate Gaussian observations, the covariance matrix S computed with denominator $N-1$ satisfies $(N-1)S \sim \text{Wishart}_m(\Sigma, N-1)$. Its expected log-determinant is

$$E[\log \det S] = \log \det \Sigma + b_m(N),$$

$$b_m(N) = \sum_{j=1}^m \psi\left(\frac{N-j}{2}\right) + m \log 2 - m \log(N-1).$$

Applying this identity to the two marginal covariance matrices and their joint covariance proves that the ordinary Gaussian information estimate has mean bias

$$\delta(N, p, q) = \frac{1}{2} [b_p(N) + b_q(N) - b_{p+q}(N)].$$

At $N=78$ with dimensions $p=4$ and $q=3$, δ is 0.08221750305196363 nats. Empirical normal-score ranks do not have the iid Gaussian/Wishart sampling law. We therefore evaluate this correction for GC rather than assuming it is exact after rank normalisation. At fixed N , p and q , δ is unchanged by a row permutation, so subtracting it before permutation centring cancels identically. Corrected-and-centred GC is the same estimator as centred GC.

S1.4 Bounds for nested profiles

A simultaneous bound on the estimation error of every pairwise mean-profile contrast also bounds the endpoint contrast by the same radius c . Because the range is the maximum pairwise contrast, its estimation error is bounded by c as well. Bootstrapping complete catalogue vectors approximates a shared within-profile radius while preserving prefix dependence. Its coverage is nominal rather than an exact finite-sample guarantee. The estimated range is at least the absolute endpoint difference. With a shared radius, a range within tolerance implies an endpoint within tolerance, while an endpoint outside tolerance implies a range outside tolerance. The joint decision therefore reduces to the range rule. A hump-shaped profile may have an out-of-tolerance range and a near-zero endpoint; these summaries are not independent confirmations.

S2 Focused implementation and study record

S2.1 Exact design and computation

The focused v0.6 configuration defines 200 independent parent catalogues per design, target and A/B stream. Each parent contains 500 observations. A uses information targets 0.1 and 0.4 nats and prefixes 78, 156, 300 and 500; independent B uses targets 0.1, 0.15, 0.4 and 0.6 with prefixes 78 and 500. The six development designs are L1, L3, T3, Q1, Q1x and Q3. H1 and H3 reserve the absolute-value geometry. Independence controls add L1, L3 and T3 at zero information, using $N=78$ and 500 without common-size subsampling.

Table S1 Generating distributions in the focused study. The signal-variance fraction s is chosen separately to match each total-information target

Design	Dimensions p and q	Active pairs	Signal or transformation
L1	1 and 1	1	Linear
L3	4 and 3	3	Linear
T3	4 and 3	3	L3 with invertible shear
Q1	1 and 1	1	Quadratic
Q1x	4 and 3	1	Quadratic plus independent coordinates
Q3	4 and 3	3	Quadratic
H1	1 and 1	1	Absolute value
H3	4 and 3	3	Absolute value

The common-size procedure uses five random subsets of 78 observations from each larger prefix, with fresh preprocessing and permutations. Subsets may overlap. Its target remains 78 even for an original (500,500) comparison. The baseline estimate is reused at $N=78$, and identical preassigned calculations are reused across E2 and E3. All k values 3, 4, 5 and 8 are computed using KSG estimator 1 [1], with centred KSG4 primary. GC [2] is retained raw, permutation-centred and analytically corrected. No score is truncated at zero and no data-dependent k -selection rule is fitted.

Table S2 Exact workload specified by v0.6. Extended calls use 500 rather than 100 permutations; their first 100 are identical. They are already included in the baseline-call count

Stage	Parents	Baseline calls	Subset calls	Extended baseline calls
Development	7,800	20,400	60,000	2,040
Reserved validation	2,400	6,400	20,000	640
Total	10,200	26,800	80,000	2,680

The bank specifies 106,800 estimator calls and 11,752,000 permutations in total. Twenty parent IDs per design, target and stream are selected independently before outcomes for the 500-permutation baseline check. Its extension does not include P1 subsets. Because the bank reuses samples across outcomes, neither its multiple prefixes nor its analysis cells are independent replications of all scientific comparisons.

Development and reserved seed bases are 202611010000 and 202612010000. The configuration records a ten-component SeedSequence key containing schema, namespace, stage, design, target, stream, replicate, random role, prefix and subset identifiers. Different streams generate observations, subsets and permutations. The frozen analysis specification separately records the bootstrap stream. Worker scheduling does not determine these keys.

To accelerate computation, the implementation reuses the joint nearest-neighbour calculation across k values. The dense backend caches within-block maximum-norm distance matrices; the tree backend uses nearest-neighbour queries. Both preserve the archived preprocessing, strict boundary counts, permutation order and definitions. Timing measurements guided backend selection. The dense backend requires storage that grows quadratically with sample size; no survey-scale performance claim is made.

S2.2 Truth and implementation validation

The unchanged truth implementation was first checked through 12 unit tests. The estimator-validation package contained 37 tests in total, including generators, estimators and random streams. Its archive regression reproduced all 400 Gaussian calibration rows at $N=78$ and correlation 0.4 exactly for GC and KSG5. Seven additional tests compared accelerated and reference implementations, including weak and strong dependence across all eight

designs, tied distances, different k orders, duplicate-vector rejection and the first-100-permutation identity within a 500-permutation run. These tests establish agreement for the tested inputs; they are not empirical calibration of the new scientific outcomes.

The focused truth registry contains 32 design–target entries. The maximum difference between its numerical reference routes is $7.46\text{e-}14$ nats and the maximum target-matching error is $1.13\text{e-}13$ nats, against a tolerance of 10^{-8} nats. Twenty preassigned generator-to-density Monte Carlo checks each use 10 million draws: eight distinct quadratic signal settings, eight absolute-value settings, and four T3 target settings. All satisfy the three-standard-error diagnostic; the largest absolute z score is 1.4183. The check is a per-setting stochastic diagnostic without adjustment across the 20 checks, and every outcome is retained.

Q1x shares its scalar information with Q1; no additional target-specific Q1x Monte Carlo run is claimed. Its independently implemented generator had already passed the representative $s=0.64$ check. L1 and L3 use analytic target information and completed representative generator checks. Zero-information controls verify all eight generators, including the within-D shear in T3. Evaluating reserved-design truth and generator consistency exposes no reserved finite-sample estimator result.

The validation outputs record Python 3.12.14, NumPy 2.3.5 and SciPy 1.17.0. The focused code fixes numerical-library thread counts per worker and records its environment. Source hashes, numerical checks and the complete truth registry are in `revision_study/validation/` within Online Resource 2. The original validated implementation remains separately in `mi_estimator_validation/`.

S2.3 Analysis rules and protocol history

Each mean error uses its assigned 200 independent catalogue replicates. Any nonfinite assigned error prevents an available-case scientific summary: the unit retains its assigned count and failure count and is classified as indeterminate. A profile with a missing prefix is not bootstrapped as complete. Ratios alone have planned positivity exclusions, reported against the full assigned count. Student-t intervals concern mean Monte Carlo errors; they are not uncertainty intervals for an individual future sample. Ranking and ratio summaries remain descriptive.

Profile uncertainty uses 9,999 resamples of whole catalogue vectors. The centred bootstrap mean-profile range is the maximum absolute centred pairwise contrast, providing an approximate shared radius for nominal 95% within-profile intervals as described in S1.4. For original-bank classifications, the tolerance is $\max(0.01 \text{ nats}, 10\% \text{ of the reference information})$. The reference is the true information for absolute calibration and profiles and I^A for comparisons; I^A is the smaller or equal truth in that original grid. The extension reports continuous errors and intervals without these tolerance categories. Classification addresses mean error within the simulated units, without simultaneous reliability over the entire grid.

Versions v0.1 through v0.5 were planning documents, not public preregistrations. Their development followed the original catalogue analyses, Gaussian calibration and quadratic-power results. The available v0.5 document was reconstructed from the earlier protocol and recorded amendments. Its broad grid proved computationally impractical. Version v0.6 is a documented feasibility amendment with a smaller fixed bank, five rather than twenty common-size subsets, baseline-only permutation extensions, both sample-size directions and prespecified cross-geometry comparisons. This changes the scientific scope explicitly; it is not a selective reduction after observing the new outcomes.

The local freeze manifest, configuration and analysis files identify what was fixed before the focused outcomes were examined. A local timestamp and file hash do not constitute public preregistration. Development interpretation is fixed before reserved-geometry performance is examined. Known legacy results and all protocol versions remain available in `protocol_history/` and `legacy/ESM_1/` within Online Resource 2. AI assistance in critique, protocol development, code generation, execution and writing is disclosed in the main manuscript.

S3 Earlier simulation analyses

S3.1 Calibration and independence tests

These earlier results motivated the focused study and are retained separately. The original calibration has four chemical and three kinematic coordinates, sample sizes 78, 156, 500 and 2000, correlations 0, 0.2, 0.4 and 0.6, two geometries, and 400 independent catalogues per cell, giving 12,800 catalogues. It compares uncorrected GC and KSG5 with known information under the Gaussian channel and its invertible deformation:

$$Y = \rho C_{1:3} + \sqrt{1 - \rho^2} \varepsilon, \quad I(C; Y) = -\frac{3}{2} \ln(1 - \rho^2).$$

$$D_1 = Y_1, \quad D_2 = Y_2 + \beta(Y_1^2 - 1), \quad D_3 = Y_3.$$

The deformation coefficient is $\beta=0$ or 1. It changes within-block geometry without changing true information. Calibration means, standard deviations, biases and RMSE are retained in `legacy/ESM_1/results/calibration_summary.csv`; all 12,800 replicates remain in `calibration_replicates.csv`.

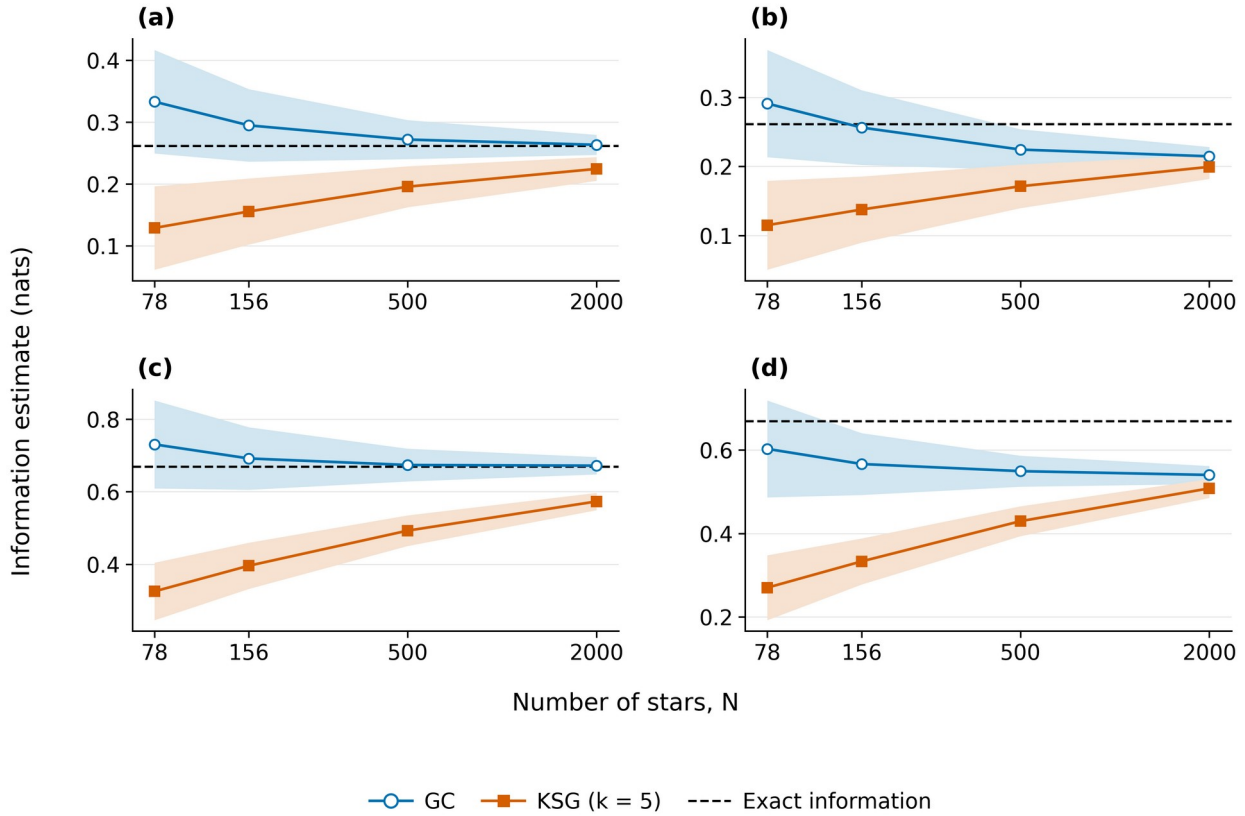


Fig. S1 Information estimates for Gaussian channels and invertibly transformed channels. GC denotes the Gaussian-copula covariance score; KSG denotes the Kraskov–Stögbauer–Grassberger estimator with neighbour order $k = 5$. Curves show means over 400 independent catalogues per point. Shaded bands show one standard deviation of the estimates, not confidence intervals for their means. Dashed lines mark the analytic information, which is unchanged by the transformation. Each catalogue has four chemical and three kinematic dimensions. Panels (a) and (b) use $\rho = 0.4$, and (c) and (d) use $\rho = 0.6$; (a) and (c) are Gaussian, while (b) and (d) apply the curved transformation

Table S3 Earlier calibration at N=78. Values are in nats; each row summarises 400 catalogues. These are raw scores, not the new permutation-centred estimates

ρ	β	Truth	GC mean	KSG5 mean	GC RMSE	KSG5 RMSE
0	0	0.0000	0.0815	-0.0014	0.0882	0.0485
0.4	0	0.2615	0.3331	0.1287	0.1100	0.1489
0.4	1	0.2615	0.2912	0.1149	0.0830	0.1601
0.6	0	0.6694	0.7305	0.3251	0.1356	0.3533
0.6	1	0.6694	0.6028	0.2699	0.1339	0.4070

The original testing grid contains 2,200 distinct catalogues. Independent Gaussian, linear ($\rho=0.35$) and one-coordinate quadratic scenarios use 200 catalogues at each of N=78, 156 and 500. Six-group scenarios add 200 catalogues at N=78 and 200 at N=156. Each grouped catalogue is tested with both global and within-group permutations; the two nulls do not constitute additional catalogues. Each test uses 199 permutations and rejects at a Monte Carlo p value no larger than 0.05, with the observed statistic included in the usual B+1 denominator [3].

The one-coordinate quadratic generator is

$$D_1 = \frac{0.8}{\sqrt{2}}(C_1^2 - 1) + 0.6\varepsilon_1.$$

Its other kinematic coordinates are independent standard normals. A separate follow-up generated 5,000 new catalogues at N=78 using the same distribution and 199 permutations. Its replicate count and new seed sequence were fixed before inspecting the follow-up outcomes, after the original 31.0% KSG5 rejection rate was known. The follow-up is not represented as a preregistered confirmation. Together, these experiments contain 7,200 current testing catalogues, separate from the 12,800 calibration catalogues.

Table S4 Earlier quadratic follow-up at N=78. Wilson intervals describe Monte Carlo rejection fractions [4]

Method	Original rate 200 catalogues	Follow-up count	Follow-up rate	95% interval
GC	5.0%	285 / 5,000	5.70%	5.09–6.38%
KSG5	31.0%	1,685 / 5,000	33.70%	32.40–35.02%
Distance correlation	32.5%	1,499 / 5,000	29.98%	28.73–31.26%
GC and KSG intersection	3.5%	170 / 5,000	3.40%	2.93–3.94%

The intersection requires both GC and KSG to reject the same null. Its rejection event is contained in either component event; reduced power is therefore expected when one component is insensitive. The alternatives were not information-matched. The linear three-pair alternative at $\rho=0.35$ has approximately 0.196017 nats, whereas the quadratic one-pair alternative has 0.4036272514 nats. The matched-information focused study addresses a different comparison.

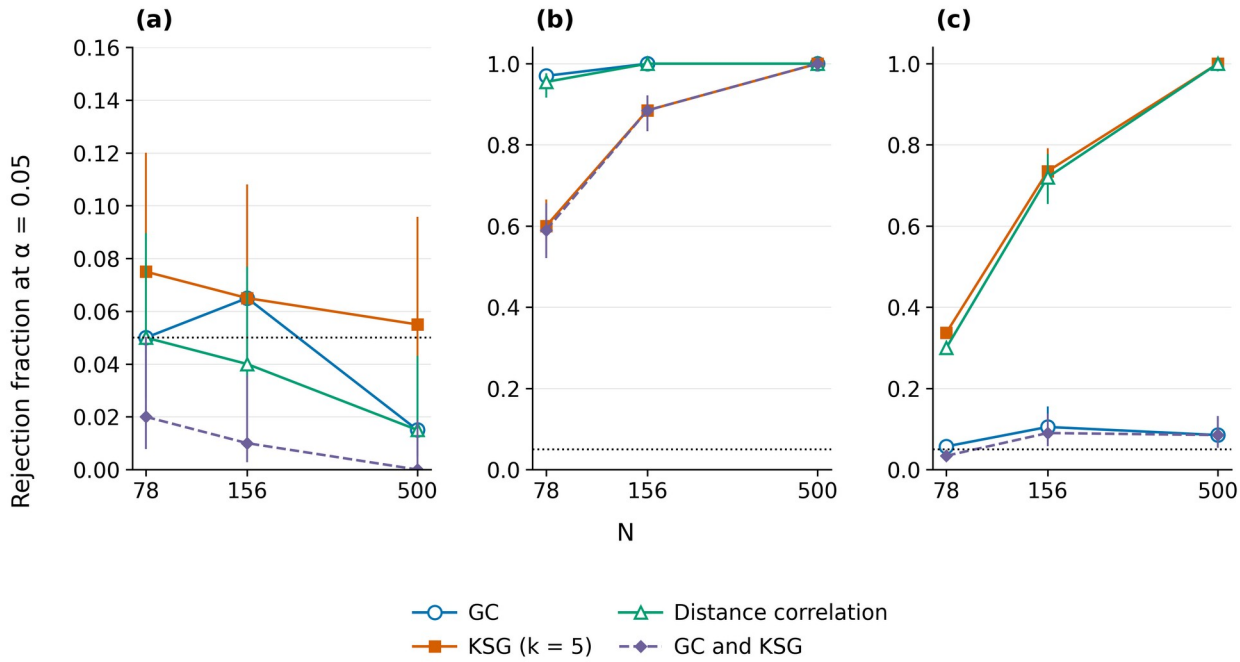


Fig. S2 Rejection fractions under independence and the specified linear and quadratic alternatives. The quadratic $N = 78$ point uses the separate 5,000-catalogue follow-up; all other points use the original 200 catalogues. Each catalogue uses 199 permutations; vertical bars are 95% Wilson intervals. The dotted line marks the nominal significance level 0.05, and panel (a) uses an expanded vertical scale. GC denotes the Gaussian-copula covariance test; KSG denotes the Kraskov–Stögbauer–Grassberger test with $k = 5$. The intersection requires both tests to reject the same null. The alternatives are not matched in mutual information. Panels (a), (b) and (c) show independence, linear dependence and quadratic dependence, respectively

S3.2 Groups and selection

The group experiment uses six nearly balanced strata. A standardised group score shifts the first two coordinates of each block by 1.2 and 0.8 times that score, with independent Gaussian residuals within groups. The domains are dependent marginally and independent within groups. At $N=78$, global rejection rates are 100% for GC, 98% for KSG5 and 100% for distance correlation; these detect genuine marginal dependence. Within-group rates are 7.0%, 8.0% and 7.5%, with Wilson intervals 4.22–11.41%, 4.98–12.60% and 4.60–12.00%. At $N=156$, the respective rates are 4.5%, 4.5% and 3.5%. The uncertainty from 200 replicates is substantial.

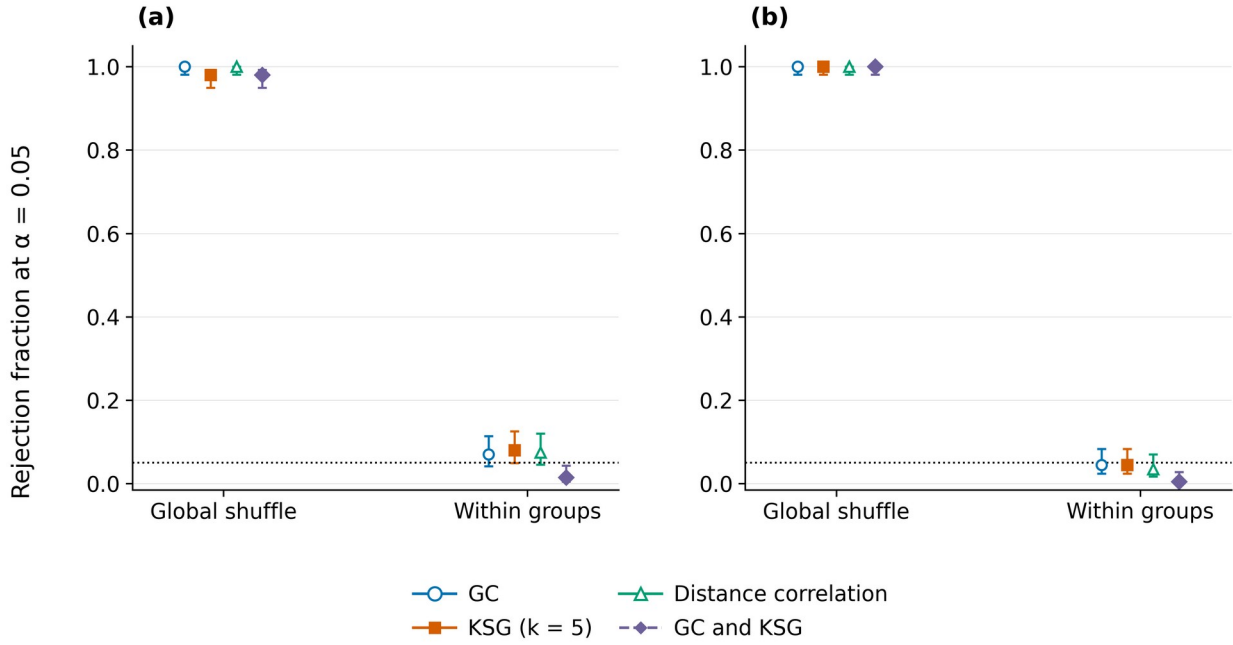


Fig. S3 Dependence generated by six group-level mean shifts. Global and within-group permutations address marginal and within-group independence, respectively. The simulated domains are dependent marginally and independent within groups. Points and bars show rejection fractions and 95% Wilson intervals over 200 catalogues per sample size. GC denotes the Gaussian-copula covariance test; KSG denotes the Kraskov–Stögbauer–Grassberger test with $k = 5$. The intersection requires both tests to reject and is especially conservative in this example. Panels (a) and (b) use $N = 78$ and $N = 156$, respectively

A separate selection illustration begins with independent scalar standard-normal variables and weights their joint distribution by

$$\pi(c, d) = \frac{1}{1 + \exp\{-\kappa(c + d - t)\}}.$$

The model uses $\kappa = 0.5, 1, 2, 4, 8$ and $t = -1.5, -1, -0.5, 0, 0.5, 1, 1.5$. Tensor-product Gauss–Hermite quadrature uses orders 80, 160, 320, 640. At $\kappa = 8$ and $t = 1.5$, the selected fraction is 0.1475, integrated mutual information is 0.3764 nats, and the Gaussian expression based on raw Pearson covariance is 0.2714 nats. The largest change in integrated information between orders 320 and 640 is 2.59×10^{-6} nats. The raw-covariance expression does not include marginal normal scoring and is not GC.

The selection function depends jointly on both domains. If the parent domains are independent, selection with weights that factorise across them preserves that independence. A kinematic cut alone therefore does not create chemical–kinematic dependence in this case. This joint selection model is illustrative; it was not fitted to the stellar catalogue or its targeting process.

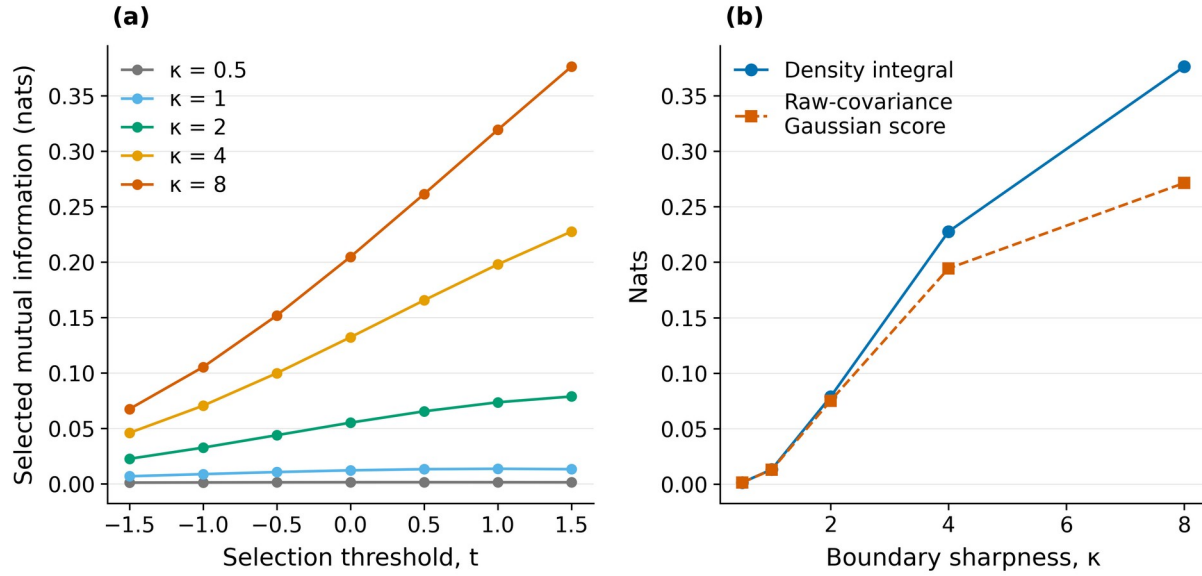


Fig. S4 Mutual information induced by joint logistic selection from independent standard-normal variables. Panel (a) integrates the selected density at five boundary sharpnesses. Panel (b) compares that integral with the Gaussian score from the raw selected covariance at $t = 1.5$. All values use order-640 Gauss–Hermite quadrature and describe population quantities. The raw-covariance score does not use marginal rank normalisation and is distinct from the Gaussian-copula statistic used for the other GC analyses. This joint-selection example does not model the catalogue’s speed cut or targeting

S3.3 Earlier three coordinate quadratic design

An earlier, separately preserved archive generated 500 catalogues at $N=78$ in each of seven scenarios. Its quadratic model has three active pairs, with each kinematic coordinate equal to a centred unit-variance squared chemical signal plus Gaussian noise of standard deviation 0.75. Rescaling each kinematic coordinate by 1.25 yields signal and noise coefficients 0.8 and 0.6, so the signal-variance fraction is again $s=0.64$. Its total information is three times the scalar value, or 1.2108817541 nats. It is therefore a different distribution from the later one-active-pair design, whose information is 0.4036272514 nats.

The earlier KSG5 test rejected in 494 of 500 catalogues (98.8%). The preserved summary is `validation/matched_regime_summary.csv` inside `legacy/ESM_1/provenance/earlier_validation_20261003.zip`. The generator is in that ZIP’s `validation/validate_matched_regime.py`. The 98.8% and 33.70% rates must not be pooled or described as a reproduction discrepancy. The earlier archive records squared distance correlation under `dCor2` and has its own sensitivity settings; it is preserved for traceability and is not used to replace the active results.

S4 Stellar catalogue and observational sensitivity

S4.1 Source verification and null definitions

The observational example uses the abundance and space-velocity tables of Nissen and Schuster [5], CDS/VizieR J/A+A/511/L10, DOI 10.26093/cds/vizie.35119010. The source export was accessed on 2 October 2026. There are 53 UVES and 47 FIES observations, including six duplicate stars. Preferring UVES for duplicates leaves 94 unique stars. All 658 adopted numerical cells and the instrument labels match the included source export. The retained sample comprises 78 stars with published local-standard-of-rest speed greater than 180 km s^{-1} .

The chemical vector contains $[\text{Fe}/\text{H}]$, $[\text{Mg}/\text{Fe}]$, $[\text{Si}/\text{Fe}]$ and $[\text{Ni}/\text{Fe}]$, and the kinematic vector contains U , V and W . Chemical population labels do not define the target. We retain the published velocities for this archival

methodological example, without a new astrometric reduction. The sample is not representative of the entire Galactic halo, and the local velocities are not orbital actions.

The principal restricted null crosses instrument with three quantile bins of published [Fe/H]. The FIES cell sizes are 12,14,15 and the UVES sizes are 16,11,10. Permutations move complete velocity rows within these cells. Instrument is a provenance variable; metallicity is an astrophysical predictor and part of the chemical vector. Preserving metallicity bins intentionally retains some real chemical–kinematic population structure. Coarse bins do not fully condition on continuous metallicity, and the resulting centred score is not conditional mutual information.

Table S5 Permutation counts in the retained observational analyses. The Monte Carlo p value denominator is B+1

Analysis	B	Denominator
Global comparison	1,000	1,001
Instrument by metallicity tertiles	2,000	2,001
Neighbour-order grid k=3,4,5,6,8,10	2,000	2,001
Each alternative null and [Fe/H] removal	1,999	2,000
Each binary-pair exclusion	1,999	2,000
Each rounding or added-noise draw	499	500

GC is the uncorrected normal-score covariance statistic in these legacy analyses. KSG uses k=5 unless a neighbour order is stated. Distance correlation is reported as the non-negative square root of the squared sample statistic [6]. For these stars it is 0.4504546771; its square is 0.2029094161. Both are dimensionless. The earlier provenance archive’s dCor2 scale must not be confused with the active distance-correlation scale.

S4.2 Observed scores and alternative nulls

Table S6 Observed statistics and permutation ensembles for the 78-star sample. GC and KSG5 scores are in nats; distance correlation is dimensionless

Null	Statistic	Observed	Null mean	Null 95th percentile	p
Global	GC	0.1483	0.0803	0.1414	0.0320
Global	KSG5	0.2106	0.0080	0.0896	0.0020
Global	Distance correlation	0.4505	0.2914	0.3391	0.0010
Instrument × thirds	GC	0.1483	0.1601	0.2277	0.5932
Instrument × thirds	KSG5	0.2106	0.0838	0.1642	0.0060
Instrument × thirds	Distance correlation	0.4505	0.3492	0.3952	0.0010

KSG5 and distance correlation remain unusual relative to the principal stratified null; GC does not. Instrument-only permutations give GC p=0.0420, whereas metallicity-tertile permutations alone give p=0.4885. In these comparisons, the large change in GC significance is mainly associated with preserving metallicity groups. This is a change in the scientific null, not direct evidence that instrumental contamination caused the global association.

Instrument crossed with halves or quartiles gives GC p=0.4675 and 0.6990, while KSG5 gives p=0.0050 and 0.0070. The smallest quartile cell contains seven stars. Removing [Fe/H] from the chemical vector while retaining the original strata gives GC p=0.5525, KSG5 p=0.0040 and distance-correlation p=0.0005. These are exploratory sensitivity analyses, not independent replications or evidence of a specific nonlinear physical mechanism.

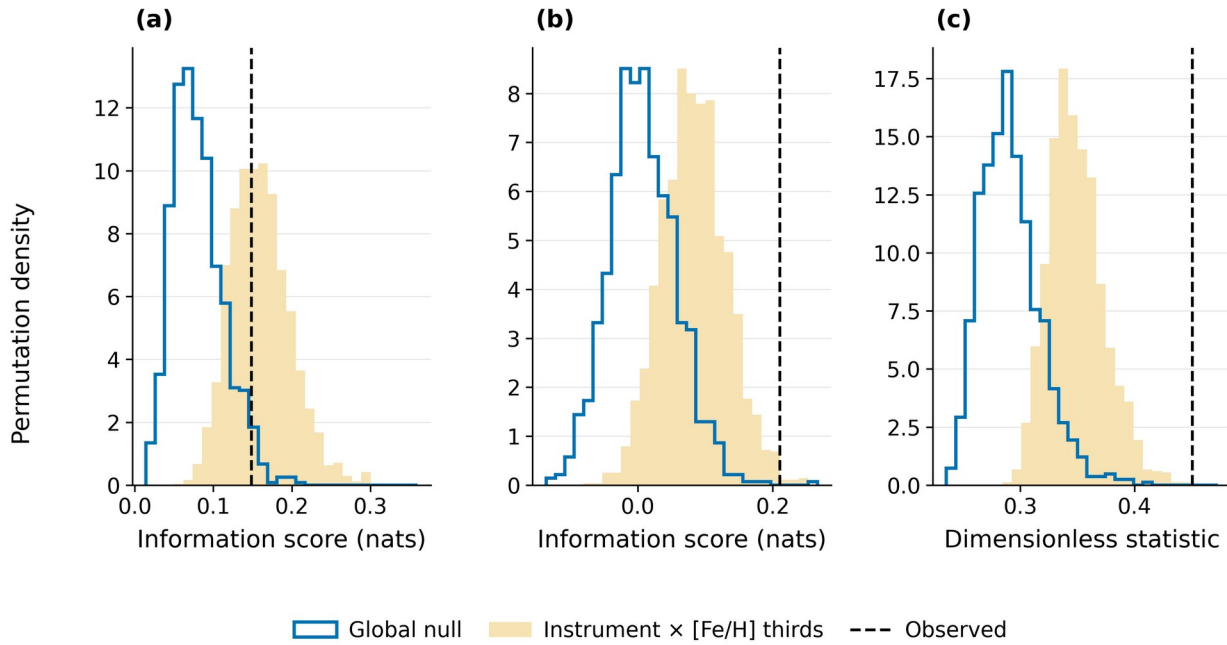


Fig. S5 Permutation distributions for the 78-star catalogue. The observed statistic is unchanged when the null changes. Global permutations use 1,000 draws; instrument-by-metallicity permutations use 2,000. Histogram densities are normalised separately. GC denotes the Gaussian-copula covariance score; KSG denotes the Kraskov–Stögbauer–Grassberger estimator with $k = 5$. Distance correlation is dimensionless, while the information scores are in nats. Coarse stratification does not estimate conditional mutual information. Panels (a), (b) and (c) show GC, KSG5 and distance correlation, respectively

S4.3 Neighbours, rounding, noise and a binary pair

Across $k=3,4,5,6,8,10$, KSG estimates range from 0.2085 to 0.3379 nats, and stratified p values range from 0.0005 to 0.0060. Holm’s step-down adjustment within this six-test family gives adjusted p values no larger than 0.0090 [7]. It does not adjust for every exploratory choice in the study. The variation in score magnitude also cautions against a precise physical interpretation of one neighbour order.

All 78 chemical and joint vectors are unique, but individual coordinates are rounded; [Ni/Fe] has only 22 distinct published values. Rounding sensitivity adds independent uniform perturbations within ± 0.005 dex to each chemical coordinate and ± 0.5 km s $^{-1}$ to each velocity. Each of 100 draws receives a newly computed 499-permutation null. Original sample membership and stratum labels remain fixed.

A separate 100-draw exercise adds independent Gaussian chemical noise with standard deviations (0.035, 0.035, 0.035, 0.010) dex and velocity noise with standard deviations (12, 16, 9) km s $^{-1}$. These approximate the precision values reported by Nissen and Schuster; they are not individual covariance matrices. Each draw uses a fresh 499-permutation null. The KSG5 score’s 5th–95th percentile range is 0.1753–0.2928 nats. Added-noise draws assess sensitivity of already measured values; they neither estimate a latent deconvolved distribution nor provide sampling confidence intervals.

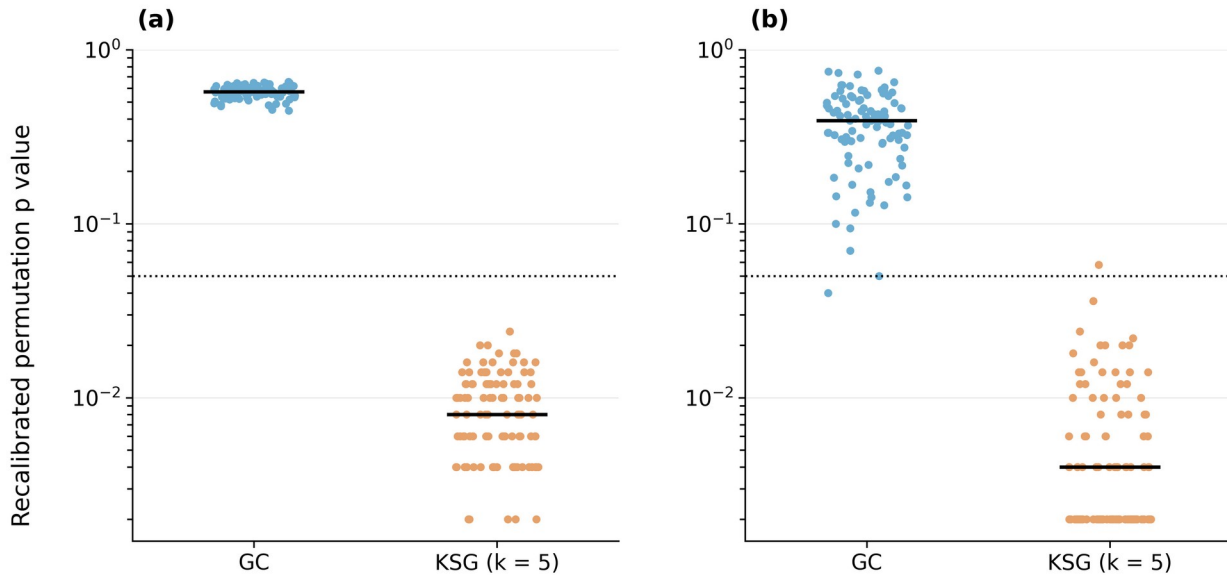
Table S7 Rounding and added-noise sensitivity. Counts are out of 100 draws; p-value ranges are empirical 5th–95th percentiles

Exercise	Method	Rejections	p-value percentile range
Publication rounding	GC	0 / 100	0.4878–0.6401
Publication rounding	KSG5	100 / 100	0.0039–0.0180
Added Gaussian noise	GC	2 / 100	0.1152–0.6292
Added Gaussian noise	KSG5	99 / 100	0.0020–0.0200

The visual binary G112-43/G112-44 can violate independent stellar sampling. Removing either component or both while preserving the original bin assignments gives the results in Table S8. KSG5 and distance correlation remain significant under these tested exclusions. This does not establish independence of every remaining star or eliminate shared calibration errors.

Table S8 Binary exclusions using fixed baseline strata and 1,999 restricted permutations

Excluded	N	GC p	KSG5 p	Distance correlation p
G112-43	77	0.6435	0.0170	0.0020
G112-44	77	0.5925	0.0150	0.0010
Both components	76	0.6450	0.0025	0.0030

**Fig. S6** Permutation p values after 100 rounding perturbations and 100 separate added-noise perturbations. Each point uses 499 instrument-by-metallicity permutations. Horizontal black marks show medians and the dotted line marks 0.05. GC denotes the Gaussian-copula covariance test; KSG denotes the Kraskov–Stögbauer–Grassberger test with $k = 5$. Original sample membership and stratum labels remain fixed. The draws assess perturbation sensitivity and are not independent replications of the stellar population. Panels (a) and (b) show publication-rounding and added-noise perturbations, respectively

The earlier provenance archive uses two sets of 200 perturbations with 199 restricted permutations and recomputes metallicity bins for binary exclusions using 2,000 permutations. The retained active analyses above instead use 100 draws with 499 permutations and fixed bins with 1,999 permutations. These differences explain why similarly named historical tables contain different values. Their source bytes remain preserved.

S5 Additional focused study results

The additional figures report the prespecified nearest-neighbour sensitivity and the reserved absolute-value geometries. H1 has one active scalar pair; H3 has three active pairs in blocks of dimensions four and three. Reserved calculations use independent random streams and the same fixed estimators, targets and analysis rules. Their performance was examined after development interpretation was recorded.

Complete results are provided in the stage-specific analysis directories under revision_study/results/ in Online Resource 2. Calibration and comparison tables include means, Monte Carlo standard errors, confidence intervals and RMSE; profile tables include the simultaneous contrast bounds. The paired common-size and permutation-extension tables retain their declared comparison units. These figures do not select a neighbour order or establish performance for all nonlinear distributions.

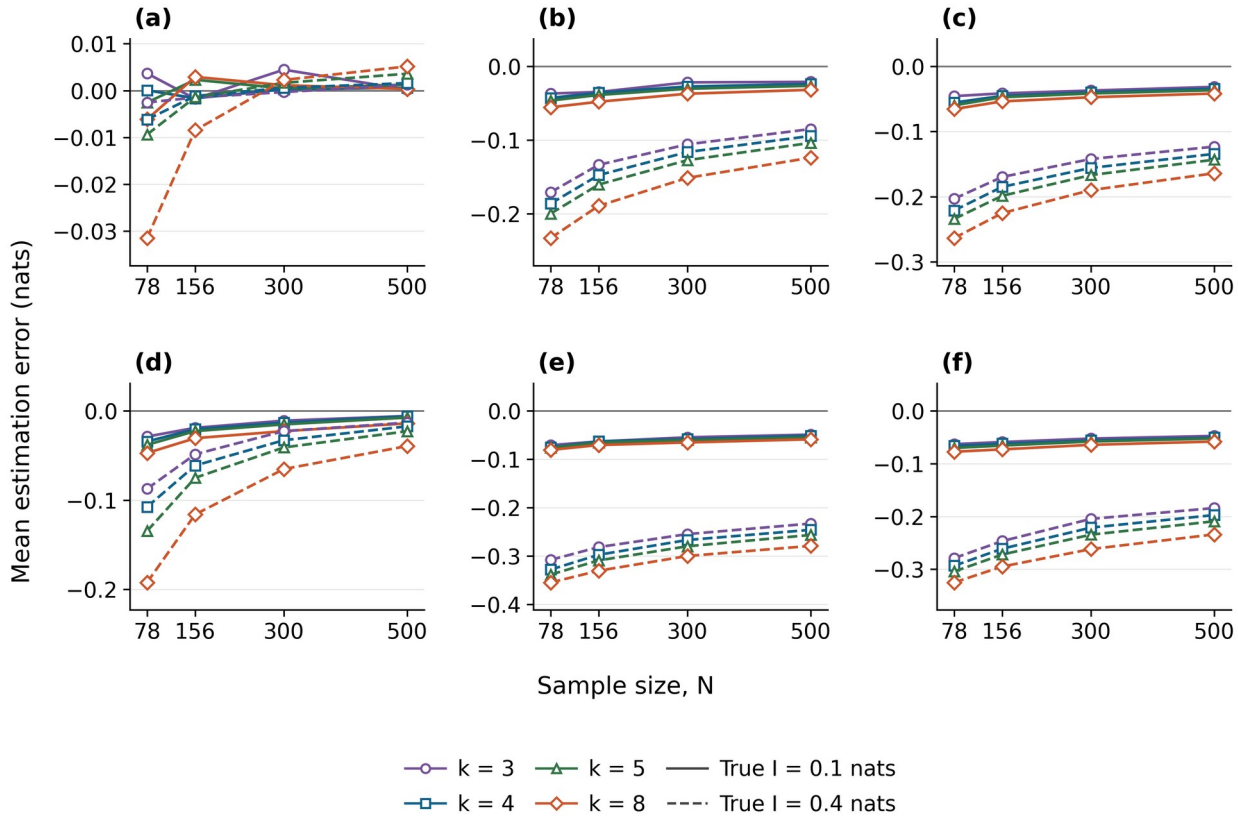


Fig. S7 Prespecified nearest-neighbour sensitivity in the six development designs. Curves show mean estimation error for permutation-centred KSG at $k=3,4,5,8$ using the same datasets and permutation sequences. Solid and dashed lines denote true information of 0.10 and 0.40 nats. Each point summarises 200 independent parent catalogues. Prefixes within a parent are nested. No k is selected from the displayed performance; full uncertainty summaries are archived. Panels (a)–(f) show L1, L3, T3, Q1, Q1x and Q3, respectively

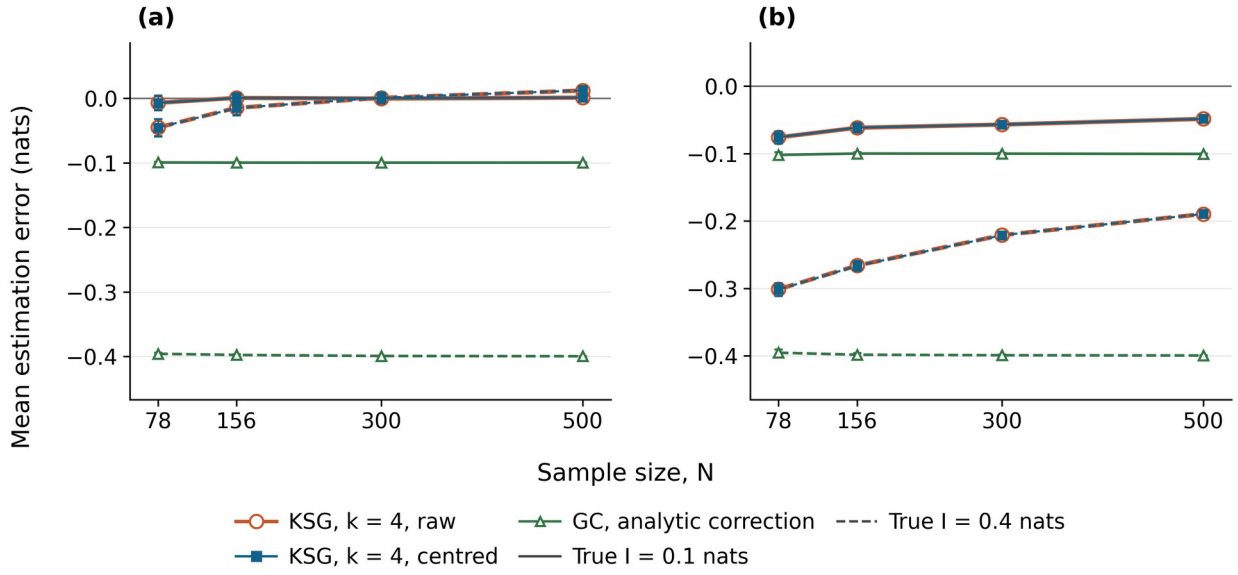


Fig. S8 Absolute calibration in the reserved H1 and H3 geometries. Mean estimation errors are shown for raw and centred KSG4 and analytically corrected GC. Solid and dashed lines denote true information of 0.10 and 0.40 nats. Bars are two-sided 95% Student-t Monte Carlo intervals over 200 independent parent catalogues. The Gaussian correction is not assumed exact after empirical rank normalisation. Under these symmetric nonlinear alternatives, GC also has population approximation error. Panels (a) and (b) show H1 and H3, respectively

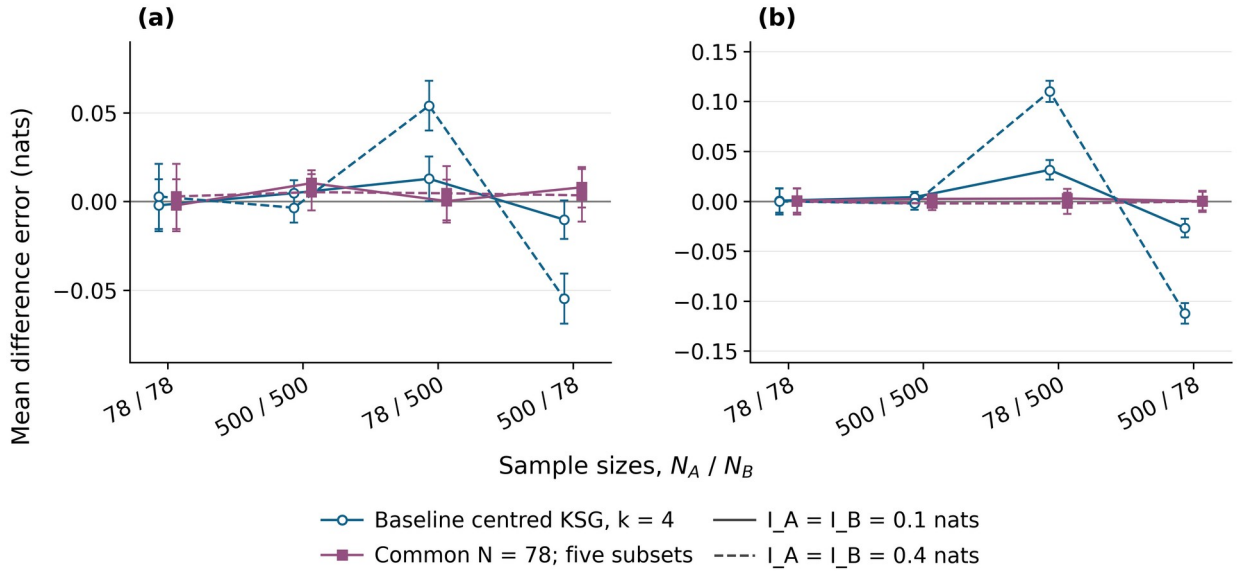


Fig. S9 Equal-information comparisons in the reserved geometries. Curves show mean errors in B-minus-A information differences for centred KSG4; the true difference is zero. Baseline estimates use the indicated sample sizes. P1 uses 78 rows on both sides, averaging five subsets from each larger source sample and reusing the baseline at $N=78$. Solid and dashed lines denote common information of 0.10 and 0.40 nats. Bars are 95% Student-t Monte Carlo intervals over 200 independent A/B pairs, reused across the specified size comparisons. Panels (a) and (b) show H1 and H3, respectively

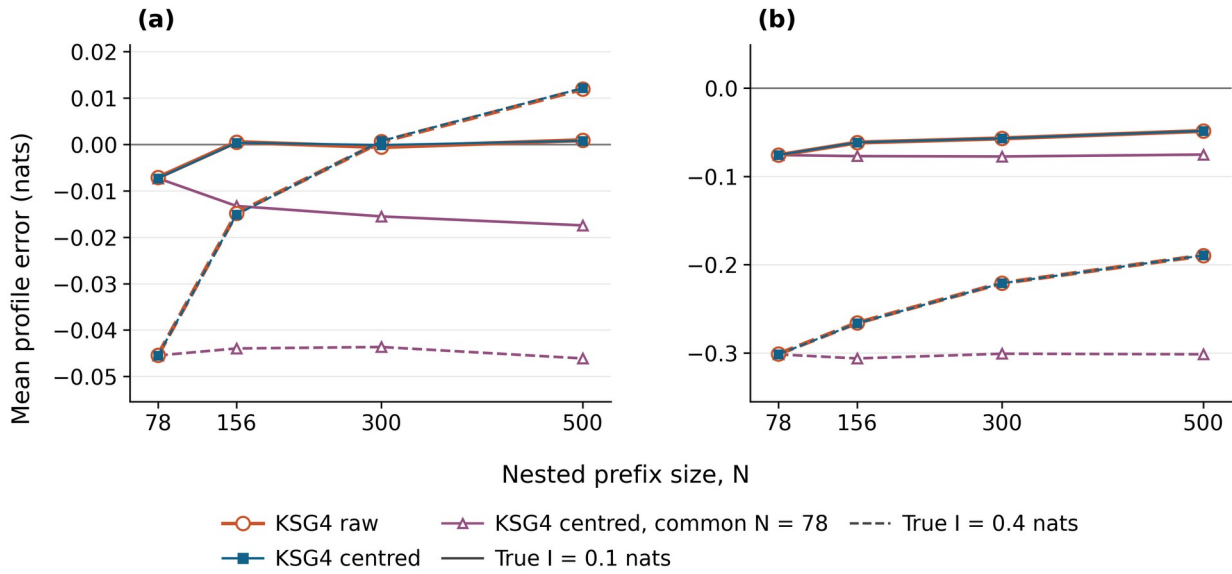


Fig. S10 Constant-information profiles in the reserved geometries. Curves show mean errors for raw KSG4, centred KSG4 and centred KSG4 with common-size subsampling. Solid and dashed lines denote true information of 0.10 and 0.40 nats. Every prefix is drawn from the same population; each curve summarises 200 independent parent catalogues. Whole-catalogue bootstrap intervals for the endpoint and mean-profile range are supplied in the archived profile table. A flat mean profile does not establish accurate absolute information. Panels (a) and (b) show H1 and H3, respectively

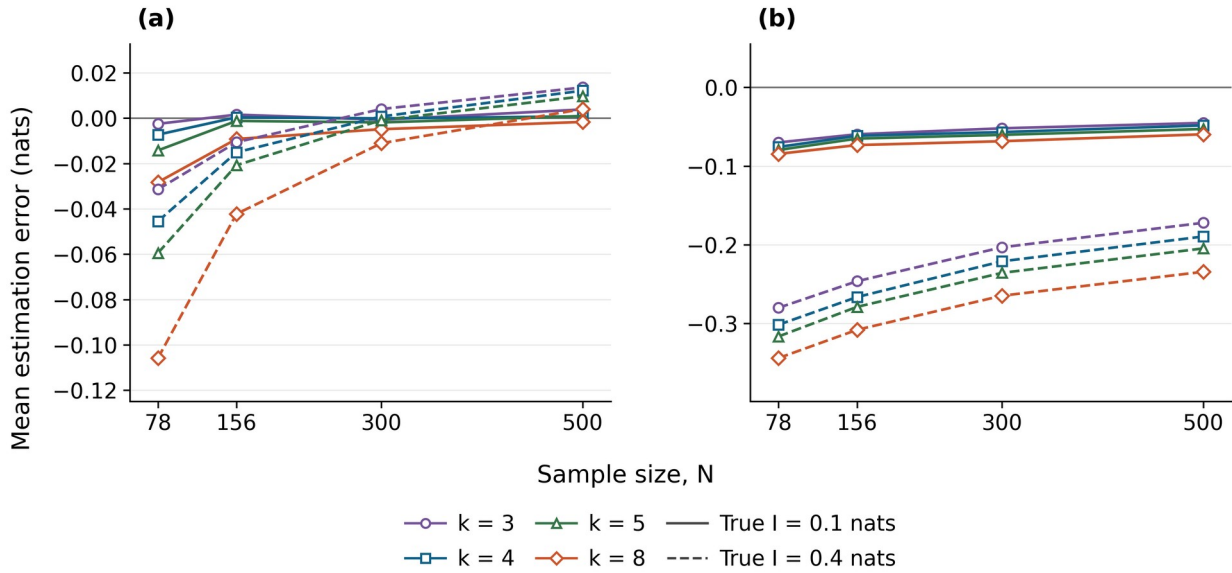


Fig. S11 Prespecified nearest-neighbour sensitivity in the reserved geometries. Mean estimation errors are shown for centred KSG at $k=3,4,5,8$. Solid and dashed lines denote true information of 0.10 and 0.40 nats. Each point summarises 200 independent parent catalogues; datasets and permutations are shared across k . These fixed sensitivity analyses assess extension to the absolute-value family without selecting a preferred k from its observed performance. Panels (a) and (b) show H1 and H3, respectively

S6 Post hoc audit of existing results

This audit was specified after the focused results had been examined. It reanalyses the existing catalogue records and observational data; it generates no new simulation catalogues. The original development and reserved-stage labels identify the source records, but the additional summaries below are descriptive post hoc analyses and are not a new confirmatory validation. The analysis script, source hashes, complete cell summaries and paired replicate outcomes are supplied in revision_audit/ within Online Resource 2.

S6.1 Raw and centred estimates

The audit pairs raw KSG4 and permutation-centred KSG4 within each catalogue and procedure, using the original 100-permutation reference. The 262 calibration cells comprise 134 baseline settings and 128 common-78 settings, including reuse of the baseline estimates when $N = 78$. Across these cells, the largest absolute difference between cell means is 0.001244 nats, the largest within-cell root mean squared correction is 0.006242 nats, and the largest absolute change in absolute-estimation RMSE is 0.000816 nats.

Across 288 comparison/procedure cells, the largest absolute shift in mean estimated difference is 0.001458 nats and the largest absolute change in comparison RMSE is 0.001840 nats. The largest absolute change in correct-ordering probability is 2.5 percentage points across the 128 cells with nonzero true differences. Across the 160 equal-information cells, where no true ordering exists, the largest change in the proportion favouring B is 3.0 percentage points. Up to 13 of 200 estimated differences change sign within a cell. These cellwise extrema are descriptive, and cells reuse catalogue records. No equivalence margin or equivalence test was prespecified.

The primary Q1x example in Table S9 uses true information values of 0.4 nats for A and 0.6 nats for B, giving a true B-minus-A difference of 0.2 nats. Centring leaves the large negative comparison error unresolved. Common-size subsampling improves comparison performance for both raw and centred scores, but substantial error remains. The full audit tables report signed, absolute and root-mean-square raw-centred differences, together with the direction of every sign change.

Table S9 Raw and centred KSG4 for the Q1x comparison at original sizes $N(A) = 500$ and $N(B) = 78$; 200 independent catalogue pairs per row; errors and RMSE are in nats; ordering intervals are 95% Wilson intervals

Procedure	KSG4 score	Mean error	RMSE	Correct ordering (95% interval)
Original sizes	Raw	-0.2749	0.2821	11.5% (7.8–16.7)
Original sizes	Centred	-0.2752	0.2824	13.0% (9.0–18.4)
Common size 78	Raw	-0.1908	0.2016	55.0% (48.1–61.7)
Common size 78	Centred	-0.1907	0.2015	55.0% (48.1–61.7)

S6.2 Ordering at equal and unequal sample sizes

Table S10 reports centred KSG4 ordering probabilities for equal and unequal original sample sizes, with no subsampling. Every cell has true information of 0.4 nats for A and 0.6 nats for B, so B has stronger dependence. Equal sample sizes should not be confused with equal information. For equal-information controls, there is no correct ordering. The complete audit therefore reports mean difference error and the fraction of estimates favouring B; these fractions are not false-positive rates for a significance test.

Table S10 Correct ordering of B above A using original-size centred KSG4 estimates; percentages with 95% Wilson intervals from 200 independent catalogue pairs in each cell; the original reserved designs H1 and H3 are retained and this additional table is post hoc

Design	N(A), N(B) 78, 78	N(A), N(B) 500, 500	N(A), N(B) 78, 500	N(A), N(B) 500, 78
L1	86.0% (80.5–90.1)	100.0% (98.1–100.0)	96.0% (92.3–98.0)	92.5% (88.0–95.4)
L3	80.0% (73.9–85.0)	99.5% (97.2–99.9)	99.5% (97.2–99.9)	53.0% (46.1–59.8)
T3	86.0% (80.5–90.1)	99.5% (97.2–99.9)	99.0% (96.4–99.7)	56.5% (49.6–63.2)
Q1	78.5% (72.3–83.6)	100.0% (98.1–100.0)	99.5% (97.2–99.9)	55.5% (48.6–62.2)
Q1x	56.0% (49.1–62.7)	86.0% (80.5–90.1)	96.0% (92.3–98.0)	13.0% (9.0–18.4)
Q3	66.5% (59.7–72.7)	95.0% (91.0–97.3)	99.5% (97.2–99.9)	24.0% (18.6–30.4)
H1	86.5% (81.1–90.6)	100.0% (98.1–100.0)	99.0% (96.4–99.7)	80.5% (74.5–85.4)
H3	69.5% (62.8–75.5)	98.5% (95.7–99.5)	100.0% (98.1–100.0)	19.0% (14.2–25.0)

S6.3 Paired improvement and mean squared error

For the primary Q1x comparison in Table S9, 26 of 200 centred original-size estimates have the correct sign, compared with 110 after common-size subsampling. Table S11 preserves the pairing by independent A/B catalogue pair. All 26 initially correct rankings remain correct; 84 initially incorrect rankings become correct. The paired improvement is 42.0 percentage points, with a 95% percentile bootstrap interval of 35.0–49.0 percentage points from 9,999 resamples of whole catalogue pairs (seed 2026101501). Both procedure outcomes are resampled together. This bootstrap operates on independent catalogue-level outcomes and does not resample the observations used by KSG.

Table S11 Paired correctness outcomes for original-size and common-size centred KSG4 in the primary Q1x comparison; counts sum to 200 independent catalogue pairs

Original-size outcome	Common size correct	Common size incorrect	Total
Correct	26	0	26
Incorrect	84	90	174
Total	110	90	200

The two-sided exact McNemar test on the 84 discordant pairs gives $p = 1.03 \times 10^{-25}$. This post hoc supporting test assesses equality of the paired correctness probabilities; it does not establish that the common-size ordering probability exceeds one half. The latter estimate is 55.0%, with a Wilson interval of 48.1–61.7%.

Using the empirical variance with denominator 200 gives the exact decomposition $MSE = \text{squared mean error} + \text{empirical variance}$. Before subsampling, the mean error is -0.275224 nats, MSE is 0.07973499 nats², squared mean error is 0.07574839 nats² and empirical variance is 0.00398660 nats². Squared mean error therefore accounts for 95.00% of empirical MSE in this setting. After subsampling it accounts for 89.57%, with mean error -0.190652 nats and RMSE 0.201452 nats. These percentages describe the observed error decomposition, not the fraction of incorrect rankings caused by bias, and are not unbiased estimates of population squared bias.

S6.4 Stellar neighbour order reconstruction

The retained 78-star example is recomputed at $k = 4$ to align with the primary simulation estimator, with $k = 5$ retained as a sensitivity analysis. The original global permutation stream uses 1,000 draws and seed 20260829; instrument-by-metallicity-tertile permutations use 2,000 draws and seed 20260929. The $k = 5$ permutation arrays reproduce the archived arrays within numerical tolerance. Raw scores and restricted-null summaries at both neighbour orders reproduce the archived k -grid. Table S12 reports the reconstructed reference means and centred scores; it is a post hoc reanalysis of existing observational data.

Table S12 Reconstructed KSG4 and KSG5 scores for the 78-star catalogue; scores and reference means are in nats; p values are unadjusted permutation probabilities using denominator $B + 1$

Permutation null	k	Raw score	Reference mean	Centred score	p
Global	4	0.2483	0.0103	0.2381	0.0020
Global	5	0.2106	0.0080	0.2026	0.0020
Instrument $\times$ metallicity tertiles	4	0.2483	0.0909	0.1575	0.0045
Instrument $\times$ metallicity tertiles	5	0.2106	0.0838	0.1268	0.0060

Changing the null leaves each observed raw score unchanged while altering its reference. Restricted-null p values are 0.0045 at $k = 4$ and 0.0060 at $k = 5$. For $k = 3, 4, 5, 6, 8$ and 10 , the unadjusted restricted-null p values are respectively 0.000500, 0.004498, 0.005997, 0.001000, 0.000500 and 0.000500. Holm adjustment of the unrounded values gives 0.002999, 0.008996, 0.008996, 0.002999, 0.002999 and 0.002999, respectively. The $k = 4$ value is fifth in ascending order; the sixth, $k = 5$, inherits the preceding adjusted maximum. Both therefore round to 0.0090. This adjustment concerns that six-test family only. As in S4, stratified centring does not estimate conditional mutual information, and these results do not validate an absolute information value or an evolutionary interpretation.

S7 Moderate imbalance and minimum-size matching

S7.1 Design and procedure

After the original v0.6 results were known, an independently seeded extension fixed L1 and Q1x, information targets 0.4, 0.5 and 0.6 nats, 601 independent A/B parent pairs per setting, sample-size anchors 100 and 200, ratios 1, 1.25, 1.5, 2 and 3, KSG $k=4$, 100 whole-row permutations and five-subset minimum-size matching. Each parent has 600 rows. Seven configured truth pairs are (0.4,0.4), (0.5,0.5), (0.6,0.6), (0.4,0.5), (0.5,0.4), (0.4,0.6) and (0.6,0.4); the full nine-pair cross-product was not used. A and B are independent within each pair. Prefixes, comparisons, anchors and overlapping subsets reuse the bank and are dependent across cells.

At unequal sizes, A is retained in full and five independently selected B subsets each contain N^A rows. Sampling is without replacement within a subset; different subsets may overlap. Each subset repeats standardisation, strict-neighbour counting and its own whole-row permutation calculation. The five estimates are averaged before subtracting A. Equal-size cells reuse their baseline exactly. This procedure combines equalising sample sizes with averaging; there is no separate experimental contrast that isolates either component. All scores are retained: raw and centred KSG4 and raw, centred and analytic GC. Negative estimates remain untruncated.

The unchanged implementation exactly matched the archived reference on extension adaptation checks at sample sizes 100, 200 and 600. All six design–target population truths were checked by FFT and adaptive real-space integration, agreeing within 2.6×10^{-14} nats. These numerical checks establish the declared reference values and tested implementation equivalence. L1 and Q1x were used in development; independently seeded extension catalogues do not constitute new held-out geometries.

S7.2 Outcomes and uncertainty

For every geometry, anchor, ratio, truth pair, score and procedure, the archived summaries report mean estimated B-minus-A difference, bias, comparison RMSE, absolute RMSE for each side, directional preference and tie frequency. For a nonzero true contrast $\Delta = I^B - I^A$, correct ordering means $(T^B - T^A)\Delta > 0$; an estimated tie is counted as incorrect. Equal-information controls use directional preference and difference error, without a correct-ordering or false-positive interpretation. Probability intervals are pointwise Wilson 95% intervals [4]. The 601-pair count was fixed for approximately four-percentage-point precision near probability 0.5 under normal-approximation planning.

Paired changes in ordering probability and comparison RMSE use 2,000 percentile-bootstrap resamples of independent parent-pair IDs. The analysis implementation shares a resample matrix across scores and metrics within each cell; a retained clarification records the departure from protocol wording about estimator/metric-specific keys. No raw observation rows or individual subsets are bootstrapped. Exact two-sided McNemar tests condition on the counts of improved and worsened orderings. All intervals and tests are pointwise, with no simultaneous coverage or multiplicity-adjusted claim. Full paired intervals for ordering and comparison-RMSE changes are in `paired_changes.csv`.

The primary unequal-size, nonzero-difference grid has 2 geometries $\times$ 2 anchors $\times$ 4 unequal ratios $\times$ 4 nonzero truth pairs = 64 centred-KSG4 cells. The equal-size settings add 28 geometry/truth-pair/anchor controls. Counts of improving or deteriorating cells describe this fixed grid, whose cells are dependent. They should not be interpreted as independent studies or as the probability of success in an arbitrary future dataset.

S7.3 Complete results and the controlled information swap

All 7,212 parent records and 180,300 estimator calls completed, each with 100 shared permutation evaluations. The comparison-vector array contains 140 cells $\times$ 601 pairs $\times$ 3 estimate vectors $\times$ 5 scores. Recalculation reproduced the stored summaries and paired bounds to numerical precision. No assigned parent or cell was omitted, and equal-size baseline and matched vectors were identical. Across the 64 primary cells, ordering increased in 27, decreased in 34 and was unchanged in three. Changes ranged from -33.6 to $+40.6$ percentage points; 18 positive and 19 negative pointwise paired intervals excluded zero. Comparison RMSE decreased in 34 and increased in 30 cells; 17 decreases and 18 increases had pointwise paired intervals excluding zero. Larger-B absolute RMSE increased in 51 cells. These counts are descriptive and share parent catalogues.

Main-text Table 3 fixes $N^A=200$, $N^B=400$ and swaps truth values 0.4 and 0.5 nats. The Q1x gain is $+23.1$ points [19.8,26.5] when the stronger dependence is in A; the reverse contrast loses 22.0 points [18.6,25.5]. L1 changes are smaller and their paired intervals include zero. Table S13 supplies paired discordant counts and RMSE changes; Table S14 retains mean comparison error and absolute-estimation costs. Neither improved ordering nor lower comparison RMSE establishes accurate absolute information.

Table S13 Paired outcomes for the controlled information swap at $N^A=200$ and $N^B=400$. Changes are matched minus baseline. RMSE-change intervals are pointwise 95% parent-pair percentile intervals in nats. McNemar p values are exact two-sided tests of equal paired correctness probabilities and are unadjusted. Every row contains 601 independent parent pairs; rows reuse banks

Design	I^A / I^B (nats)	Improved / worsened	McNemar p	Comparison RMSE change [95% interval]
L1	0.4 / 0.5	12 / 18	0.362	-0.0010 [-0.0035, +0.0014]
L1	0.5 / 0.4	15 / 23	0.256	+0.0011 [-0.0012, +0.0034]
Q1x	0.4 / 0.5	3 / 135	2.51e-36	+0.0274 [+0.0251, +0.0298]
Q1x	0.5 / 0.4	142 / 3	2.28e-38	-0.0331 [-0.0351, -0.0311]

Table S14 Comparison bias and absolute error for the same controlled swap. Values are in nats. A is unchanged by matching, so its absolute RMSE is identical within each pair of procedures. Five-subset matching can improve ordering while worsening B's absolute RMSE

Design	I ^A / I ^B	Procedure	Comparison bias	Absolute RMSE A	Absolute RMSE B
L1	0.4 / 0.5	Baseline	+0.0021	0.0681	0.0502
L1	0.4 / 0.5	Five subsets	−0.0001	0.0681	0.0508
L1	0.5 / 0.4	Baseline	−0.0089	0.0713	0.0483
L1	0.5 / 0.4	Five subsets	−0.0103	0.0713	0.0495
Q1x	0.4 / 0.5	Baseline	−0.0461	0.2876	0.3320
Q1x	0.4 / 0.5	Five subsets	−0.0861	0.2876	0.3713
Q1x	0.5 / 0.4	Baseline	+0.1171	0.3707	0.2534
Q1x	0.5 / 0.4	Five subsets	+0.0819	0.3707	0.2876

Within Q1x all 16 unequal cells with stronger dependence in A had higher ordering point estimates after matching and all 16 with stronger dependence in B had lower ones. This conditional pattern is not a universal matching rule. L1 changes were small and mixed. The full-grid map appears in main-text Fig. 5; plotted values include each cell's paired interval and the baseline and matched Wilson intervals. Full score-family results and equal-information controls remain in the archive. The largest extension raw-centred KSG ordering difference was 2.0 points and the largest comparison-RMSE difference was 0.00094 nats; mean permutation corrections ranged from −0.0009 to +0.0005 nats.

S7.4 Chronology and reproducibility

The runner, configuration, protocol and dependencies were hash-frozen at 10:57:01 UTC on 15 October 2026; execution began at 10:57:02. Analysis code was separately frozen at 10:59:12. Execution completed at 11:28:50 and the first recorded outcome inspection was at 11:29:34. A count erratum and the bootstrap-sharing clarification are retained with the immutable records. These local records document the declared sequence, not public preregistration or independent proof about unrecorded inspection. They do not change the original H1/H3 reserved-stage status.

The Zenodo archive (Online Resource 2) [8] retains config.json, PROTOCOL.txt and PROTOCOL_READABLE.txt, the source snapshot and seed specifications, comparison_vectors.npz, summary_all_cells.csv, paired_changes.csv, raw_centred_corrections.csv and the frozen validation records under submission_revision/extension/. The compact archive retains the complete procedure-level comparison vectors; individual parent and estimator-call checkpoints, including subset identities, can be regenerated using the frozen runner, configuration and deterministic seed rules. The archive README distinguishes output verification from full simulation regeneration. The jstp_submission/figures/ folder supplies the current publication plotting scripts, exact inputs and plotted values; reconciled_revision/figure5/ retains the earlier complete-grid build. No additional simulation catalogues were generated while preparing the submission files. The jstp_submission/ folder also supplies the numerical verification and audit reports. These reproduction checks are distinct from external statistical review or observational validation. The archived submission record predates the addition of the Zenodo DOI to the manuscript and this supplement.

References

- [1] Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. *Physical Review E* 69:066138. <https://doi.org/10.1103/PhysRevE.69.066138>
- [2] Ince RAA, Giordano BL, Kayser C, Rousselet GA, Gross J, Schyns PG (2017) A statistical framework for neuroimaging data analysis based on mutual information estimated via a Gaussian copula. *Human Brain Mapping* 38:1541–1573. <https://doi.org/10.1002/hbm.23471>

- [3] Phipson B, Smyth GK (2010) Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology* 9:Article 39. <https://doi.org/10.2202/1544-6115.1585>
- [4] Wilson EB (1927) Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22:209–212. <https://doi.org/10.1080/01621459.1927.10502953>
- [5] Nissen PE, Schuster WJ (2010) Two distinct halo populations in the solar neighborhood. Evidence from stellar abundance ratios and kinematics. *Astronomy & Astrophysics* 511:L10. <https://doi.org/10.1051/0004-6361/200913877>
- [6] Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. *Annals of Statistics* 35:2769–2794. <https://doi.org/10.1214/009053607000000505>
- [7] Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2):65–70. <https://www.jstor.org/stable/4615733>
- [8] Barua N, Oda K (2026) Comparing mutual information across unequal samples: data, code and reproducibility records. Zenodo [data set and code]. <https://doi.org/10.5281/zenodo.23466586>