

Comparing mutual information across unequal samples

Nick Barua

s.nick.barua@gmail.com

Shiga University of Medical Science <https://orcid.org/0000-0003-4641-0112>

Koji Oda

Be Bridge LLC

Research Article

Keywords: Mutual information · Finite-sample bias, Nearest-neighbour estimation, Sample-size matching, Monte Carlo simulation, Comparison accuracy

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11448808/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Comparing mutual information across unequal samples

Nick Barua^{1,2} and Koji Oda³

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

2 AN Holdings Co., 2-18 Wakamatsu-cho, Nishinomiya, Hyogo 662-0035, Japan

3 Be Bridge LLC, 3-203 Hagiwaradai Nishi, Kawanishi, Hyogo 666-0006, Japan

Corresponding author: Nick Barua · s.nick.barua@gmail.com · ORCID 0000-0003-4641-0112

Acknowledgements. This research used the Vizier catalogue access tool, CDS, Strasbourg, France. We thank Nissen and Schuster for making the abundance and velocity measurements publicly available.

Abstract

Finite-sample bias can distort comparisons of mutual information when sample sizes differ. We assess correct ordering, error in the estimated difference and the costs of matching sample sizes using an original bank of 10,200 independent parent catalogues and an independently seeded extension of 7,212 parents. The original benchmark covers linear, transformed, quadratic and reserved absolute-value dependence at sample sizes 78–500. The extension tests two geometries, anchors of 100 and 200 observations, size ratios up to three, equal-information controls and both placements of stronger dependence. In an embedded quadratic design with samples of 200 and 400 observations and a true gap of 0.1 nats, five-subset minimum-size matching increases correct ordering from 40.8% to 63.9% when the stronger dependence is in the smaller sample. Swapping the information values changes ordering from 82.7% to 60.7%. Across 64 unequal-size extension comparisons with nonzero differences, ordering point estimates increase in 27 cells and decrease in 34. Comparison root mean squared error decreases in 34 cells and increases in 30; larger-sample absolute-estimation root mean squared error increases in 51. These are descriptive summaries of dependent cells. Global permutation centring changes mean nearest-neighbour estimates little and leaves substantial comparison bias. Matching can therefore improve or impair comparison accuracy, with costs to absolute calibration. The findings apply to the tested procedure and do not establish the reliability of an unknown real-data comparison.

Keywords: Mutual information · Finite-sample bias · Nearest-neighbour estimation · Sample-size matching · Monte Carlo simulation · Comparison accuracy

1 Introduction

Mutual information is often compared across populations, observational subsets and successive prefixes of a growing sample. These comparisons require both the sign and magnitude of an estimated difference to be accurate. Bias can vary with sample size and distributional geometry, so matching sample sizes may change both. Its effects on ordering probability, difference error and absolute-estimation error must therefore be assessed separately.

Finite-sample errors in information estimation are well established [1, 2]. Nearest-neighbour estimators avoid explicit density binning, but their errors can still depend on dimension and local geometry. For the Kraskov–Stögbauer–Grassberger estimator (KSG), detecting dependence with an associated permutation test does not guarantee small magnitude error [3]. The Gaussian-copula covariance score (GC) has a further limitation: its population target can differ from mutual information outside the Gaussian-copula model [4]. This model approximation is distinct from finite-sample error.

Earlier benchmarks compare mutual-information estimators across sample sizes and noise levels [5], and recent benchmarks cover known-information distributions with sparse, high-dimensional and heavy-tailed dependence [6]. Gao, Ver Steeg and Galstyan [7] also assess ranking stability in World Health Organization (WHO) indicator data by withholding observations, recalculating pairwise information and comparing the resulting rankings with full-data rankings using Spearman correlation. In that WHO analysis, the full-data ranking is a stability reference rather than known population truth.

Holmes and Nemenman [8] develop diagnostics, uncertainty estimates and parameter-selection guidance for nearest-neighbour estimation. In neuronal data, Sheintuch et al. [9] demonstrate shuffle-subtraction overcorrection

and discuss sample-size matching. We study comparisons using independent sample pairs with controlled population information. By swapping which population has stronger dependence between fixed smaller and larger samples, we measure the consequences for ordering probability, difference error and absolute error. The contribution is a controlled comparison and a complete grid of benefits and harms for a specified five-subset matching procedure.

Galactic archaeology provides an application context. Chemical population assignments can be assessed against orbital actions and integrals of motion withheld during classification, as in the preprint study of 468,000 Milky Way stars by Barua [10]. The preprint Stellar Halo Memory by Pandey and Mondal [11] constructs permutation-centred nearest-neighbour profiles from cumulative stellar samples, whose size grows with radius. This motivates a constant-truth prefix experiment, but our synthetic designs neither reconstruct that survey analysis nor test its astrophysical conclusions. A separate archival example uses abundances and velocities from Nissen and Schuster [12, 13] to distinguish a centred magnitude from evidence against a specified null.

The original matched-information benchmark uses Gaussian, invertibly transformed, quadratic and absolute-value dependence. Single-pair and three-pair alternatives distinguish total information from its allocation across coordinates. Equal-information cross-geometry controls retain the same block dimensions, and nested prefixes isolate sample-size drift under a constant population. A separate extension evaluates ordinary sample-size imbalances at two absolute-size anchors and matches to the smaller sample. It includes equal-information controls and both placements of the stronger dependence. The two matching procedures are specified separately, and neither is assumed to correct finite-sample bias.

Population truths are independently checked, and complete replicate-level outputs are supplied. Following Morris et al. [14], we report aims, generating mechanisms, estimands, methods and performance measures separately. The records also distinguish original development, reserved-geometry validation, retrospective audits and the subsequent independently seeded extension. Local prospective specifications and their amendments are retained; the study was not publicly preregistered.

2 Methods

2.1 Estimands and permutation centring

For continuous vectors C and D with joint distribution P , mutual information is [15]

$$I(P) = E_P[\log \frac{p(C,D)}{p_C(C)p_D(D)}]. \quad (1)$$

All logarithms are natural and information is reported in nats. For a sample S of N observations, let T denote an information statistic. Independent uniform permutations move entire D rows, retaining within-block dependence. The centred statistic is

$$T_{N,B}^c = T_N(S) - \frac{1}{B} \sum_{b=1}^B T_N(S^{\pi_b}). \quad (2)$$

Permutation centring subtracts the reference mean for the observed sample. Write $a(N,P)$ for the sampling mean of the original statistic and $c(N,P)$ for that of its conditional permutation reference. The centred mean error is $a(N,P) - c(N,P) - I(P)$. With independent and identically distributed observations and independent blocks, exchangeability makes this error zero. Under dependence, however, a shuffled finite catalogue is not a fresh independent sample from the product population, and the same argument does not remove the error. The mean error in a between-population comparison is the difference between the two centred biases. Online Resource 1, Section S1, gives the identities and expectation conditions. Permutation tests address a separate null hypothesis; their validity depends on exchangeability under that null [16].

2.2 Estimators

Nearest-neighbour entropy estimation dates back to Kozachenko and Leonenko [17]. We use KSG estimator 1 with the maximum norm [3]. Each coordinate is standardised by its sample mean and sample standard deviation within

the current sample or subset. The two marginal counts include only neighbours strictly inside the k th joint-neighbour distance and exclude observation i :

$$K_k = \psi(k) + \psi(N) - \frac{1}{N} \sum_{i=1}^N [\psi(n_{C,i} + 1) + \psi(n_{D,i} + 1)]. \quad (3)$$

KSG with $k=4$ is denoted KSG4; its permutation-centred form is primary. The prespecified $k=3,5,8$ analyses are sensitivities; no k -selection rule is fitted. Strict counting moves the radius one floating-point step towards zero. Negative estimates are retained.

The Gaussian-copula score maps marginal mid-ranks to normal scores, $z_{ij} = \Phi^{-1}[(\text{rank}_{ij} - 0.5)/N]$, then computes the covariance score [4]

$$G = \frac{1}{2} \log \frac{\det(S_C) \det(S_D)}{\det(S_{CD})}. \quad (4)$$

The denominator uses the covariance matrix of the concatenated vector (C,D) ; the numerator uses its two marginal diagonal blocks. We examine G , permutation-centred G , and the analytically adjusted score $G^a = G - \delta(N,p,q)$, where

$$\delta(N, p, q) = \frac{1}{2} [b_p(N) + b_q(N) - b_{p+q}(N)], \quad (5)$$

$$b_m(N) = \sum_{j=1}^m \psi\left(\frac{N-j}{2}\right) + m \log 2 - m \log(N-1). \quad (6)$$

The correction is exact for ordinary Gaussian sample-covariance log-determinants, not generally after empirical rank normalisation. GC also has population approximation error outside its model. Centring G^a equals centring G because δ is constant across permutations, so these are not counted separately. All eleven estimator variants share 100 whole-row permutations within each call.

2.3 Known-information distributions

Independent standard-normal coordinates generate C and ε . For m active pairs,

$$D_j = \sqrt{s} g(C_j) + \sqrt{1-s} \varepsilon_j, \quad j = 1, \dots, m, \quad (7)$$

with independent noise in unused coordinates. The three unit-variance signals are

$$g_L(c) = c, \quad g_Q(c) = \frac{c^2 - 1}{\sqrt{2}}, \quad g_H(c) = \frac{|c| - \sqrt{\frac{2}{\pi}}}{\sqrt{1 - \frac{2}{\pi}}}. \quad (8)$$

Table 1 specifies the dimensions and active pairs. T3 applies the invertible shear $D_2 \leftarrow D_2 + D_1^2 - 1$ to L3, preserving mutual information. H1 and H3 reserve the absolute-value geometry for validation. Information adds across independent active pairs. Linear truth is $-(m/2) \log(1-s)$, including T3. For quadratic and absolute-value signals, $I(C;D) = m \{h(D_1) - \frac{1}{2} \log[2\pi e(1-s)]\}$. We solve for s at each target using independent characteristic-function fast Fourier transform and adaptive real-space integration, requiring agreement within 10^{-8} nats. Generator-to-density Monte Carlo checks verify the implementation separately.

Table 1 Generating distributions. The signal-variance fraction s is matched separately to the target total information

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

2.4 Simulation banks and matching procedures

The original bank has 200 independent parent catalogues of 500 observations per setting. Development uses L1, L3, T3, Q1, Q1x and Q3. Stream A targets $I=0.1$ and 0.4 nats and supplies nested prefixes $N=78,156,300,500$. Independent stream B targets $I=0.1,0.15,0.4,0.6$ and supplies prefixes $N=78,500$. Independence controls use L1, L3 and T3 at $I=0$ and $N=78,500$, without subsampling. H1 and H3 use the same non-null grid after the development interpretation is recorded. Prefixes and derived analyses are not additional independent replicates.

Absolute calibration assesses single-sample error using stream A. Comparison accuracy uses independent A/B catalogues, with information ratios I^B/I^A of 1 or 1.5 and size pairs (78,78), (500,500), (78,500) and (500,78). Equal-information cross-geometry comparisons pair $L3^A$ with $T3^B$ and $Q1x^A$ with $Q3^B$ at the same sizes. Profile accuracy uses each A catalogue's four nested prefixes, whose population information is constant. The archived protocol labels these outcomes E1, E2 and E3, respectively.

The common-size procedure, denoted common-78, targets $n_0=78$ throughout, including the (500,500) comparison. From each larger prefix, five independently selected subsets of 78 rows are drawn without replacement within each subset; subsets may overlap. Every subset receives fresh preprocessing and permutations, and their estimates are averaged. The original estimate is reused at $N=78$. Identical preassigned calculations are reused across comparisons and profiles. Common size does not equate biases of different joint distributions, and averaging subsets retains parent-sample uncertainty (Online Resource 1, Section S1). The archive denotes this fixed procedure by P1; it is not matching at the smaller original sample size in every comparison.

The subsequent moderate-imbalance extension uses independently seeded 600-row parents from L1 and Q1x at information targets 0.4, 0.5 and 0.6 nats. Each design, target and A/B stream contains 601 independent parents, giving 7,212 parents. The smaller A sample has 100 or 200 observations; B has 1, 1.25, 1.5, 2 or 3 times as many. The seven information pairs (I^A, I^B) 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). Thus both placements of a 0.1- or 0.2-nat difference are tested. Each fixed comparison uses 601 independent A/B pairs; reuse of targets, prefixes and anchors creates dependence between cells.

Five-subset minimum-size matching retains A in full and averages estimates from five independently selected B subsets, each containing N^A rows. Sampling is without replacement within each subset; subsets may overlap. Standardisation and the 100 whole-row permutations are repeated within every subset. At equal sizes the baseline estimates are reused exactly. The procedure combines sample-size equalisation with averaging; this study does not isolate their separate effects. Raw and centred KSG4 and raw, centred and analytically adjusted GC scores are retained. The extension uses previously studied geometries and supplies new catalogue-level evaluation rather than a new held-out geometry test.

2.5 Outcomes and Monte Carlo uncertainty

Absolute calibration uses error $T-I$; comparison accuracy uses $e=(T^B-T^A)-(I^B-I^A)$. Means, standard deviations and root mean squared errors (RMSE) summarise the 200 independent replicate errors in each original-bank cell. Two-sided 95% Student-t intervals describe uncertainty in each mean. For original-bank classifications, the reporting tolerance τ is the greater of 0.01 nats and 10% of the reference information. The reference is the true information for absolute calibration and profiles, and I^A for comparisons, where I^A is the smaller or equal truth in that grid. A mean is within tolerance when its whole interval lies inside $[-\tau, \tau]$, outside when the interval lies wholly beyond either boundary, and otherwise indeterminate. This classifies mean error, not accuracy of individual comparisons. The extension reports continuous errors and intervals without these tolerance classifications.

For a nonzero true contrast $\Delta=I^B-I^A$, a replicate is correctly ordered when $(T^B-T^A)\Delta>0$; an estimated tie is counted as incorrect. Correct-ordering probabilities and pointwise Wilson intervals [18, 19] describe this event. For equal-information pairs we report difference error and the direction of the estimate, without a correct-ordering or false-positive interpretation. Ratios require both estimates to be positive and are secondary to signed differences. Procedures remain paired by parent replicate; their error changes retain this pairing.

For profiles, let μ_j be the mean error at prefix j . We report every μ_j , the endpoint contrast $\Delta=\mu_4-\mu_1$, and range $H=\max_j\mu_j-\min_j\mu_j$. Resampling whole catalogue vectors 9,999 times preserves prefix dependence. The 95th percentile of the range of centred bootstrap mean errors supplies a bootstrap approximation to a shared within-profile contrast radius c , giving nominal 95% intervals $\hat{\Delta}\pm c$ and $[\max(0, \hat{H}-c), \hat{H}+c]$. A profile is within tolerance when the range's upper bound is at most τ , outside when its lower bound exceeds τ , and otherwise indeterminate. Because $\hat{H}\geq|\hat{\Delta}|$, this equals the protocol's joint endpoint-and-range rule; the endpoint reports direction. Intervals do not control multiplicity across the full grid. A flat but biased profile can have negligible drift while remaining biased in magnitude. Units with nonfinite assigned estimates are indeterminate.

Additional analyses of the completed catalogue bank compare raw and centred KSG4 estimates using paired differences, absolute and comparison root mean squared errors, and ordering probabilities. For the highlighted Q1x comparison, a percentile interval for the change in correct-ordering probability uses 9,999 resamples of independent A/B parent pairs, retaining both procedures together. The empirical mean squared error is decomposed into squared mean error and variance with denominator 200. These analyses and the added stellar $k=4$ reconstruction are post hoc; they do not change the original reserved-stage status (Online Resource 1, Section S6).

For the extension, 601 pairs were fixed for approximately four-percentage-point pointwise 95% precision near ordering probability 0.5 under normal-approximation planning [20]; reported probability intervals use Wilson's convention. Matching-minus-baseline changes in ordering probability and comparison RMSE use 2,000 percentile-bootstrap resamples of independent parent-pair IDs. The frozen analysis shares the resample matrix across scores and metrics within each cell. Exact McNemar tests use the counts that improve or worsen. No raw observations or individual subsets are bootstrapped. All intervals and tests are pointwise. Counts of improvement or harm across the 64 unequal-size, nonzero-difference KSG4 cells are descriptive summaries of dependent cells, without multiplicity adjustment. Equal-information controls report directional preference and difference error, with no correct-ordering outcome.

2.6 Computational record and study status

The design was recorded locally before its outcomes, following a feasibility amendment informed by earlier results; it was not publicly preregistered. Development interpretation preceded the reserved H1/H3 stage. Twenty catalogues per design, target and stream were preselected for a baseline sensitivity increasing the permutation count from 100 to 500, retaining the first 100 permutations. Common-size estimates were not extended. Separate streams govern data, subsets, permutations and bootstrap resampling. Full protocol labels, amendments, seeds, configuration and failures are retained in Online Resource 2.

The extension protocol, configuration, runner and dependencies were locally hash-frozen before execution. Analysis code was frozen separately after execution began and before the recorded first inspection of extension outcomes. A clarification in the archive explains that bootstrap draws were shared within cells, although the protocol wording assigned separate estimator/metric keys. The frozen scientific files are unchanged. Online

Resource 1, Section S7, gives the chronology, precision target and scope; Online Resource 2 contains the frozen records and procedure-level outputs. These local records are not public preregistration.

The implementation caches transformations and marginal distances within each current sample, shares neighbour calculations across k values and prepares every subset independently. The optimized and original implementations give equal raw and permutation scores on the fixed equivalence tests, including tied distances and the shared first 100 permutations of a 500-permutation call. Storage and timing limitations are documented in Online Resource 1. Computation used Python 3.12, NumPy 2.3.5 and SciPy 1.17.0 with one numerical-library thread per worker [21, 22]; tabulation and figures used pandas and Matplotlib [23, 24].

2.7 AI assistance

ChatGPT/Codex (OpenAI) and Claude (Anthropic) assisted with literature searches, methods and protocol development, coding, simulations and validation, figures, and manuscript drafting and editing. The authors are responsible for the methods, results and final text.

3 Results

3.1 Verification and computational record

The two numerical truth calculations agreed to at most 7.46×10^{-14} nats across the 32 design–target entries, with target-matching error at most 1.13×10^{-13} nats. All 20 preassigned generator-to-density checks, each using 10 million draws, met the three-standard-error diagnostic; the largest absolute z score was 1.42. The optimized estimator matched the reference implementation on the fixed equivalence checks. Online Resource 1 documents these checks separately from the scientific outcomes.

All 10,200 original parent catalogues completed without failures: 7,800 in development and 2,400 in the reserved stage. The original bank comprises 106,800 baseline-plus-subset estimator calls and 11,752,000 permutation evaluations. Development interpretation preceded generation of the reserved catalogues; settings were not retuned. The retrospective audit reused that bank. The subsequent extension completed all 7,212 additional parents and 180,300 estimator calls, with 18,030,000 shared permutation evaluations and no omitted parent or prespecified cell. Equal-size matching reproduced its baseline vectors exactly. Across the two banks there are 17,412 distinct parent catalogues; the per-comparison independent counts remain 200 and 601, respectively.

3.2 Absolute calibration

Permutation-centred KSG4 had small mean values under the six independence controls: -0.0034 to 0.0023 nats, with every 95% Monte Carlo interval including zero. Under dependence, calibration varied substantially by dimension and geometry (Fig. 1). At true information 0.4 nats, Q1 had a mean of 0.3826 at $N=500$. Embedding the single quadratic pair in dimensions (4,3), as Q1x, reduced that mean to 0.1543. At $N=78$, the Q1x mean was 0.0722, giving an error of -0.3278 nats (95% interval -0.3364 to -0.3192).

The analytically adjusted GC score behaved differently. At truth 0.4 nats, its means for L1 and L3 at $N=78$ and 500 lay between 0.3908 and 0.3968, all within the reporting tolerance. For the information-preserving T3 shear, means were approximately 0.315–0.320 and outside tolerance. Quadratic designs gave means near zero. These patterns separate a finite-sample correction from the population covariance score’s inability to represent some nonlinear dependence.

Raw and centred KSG4 behaved similarly in the additional audit. Across 262 estimator cells, their largest absolute mean difference was 0.00124 nats and the largest within-cell root mean squared change was 0.00624 nats. Absolute-estimation RMSE changed by at most 0.00082 nats. Across 288 comparison/procedure cells, the largest mean difference change was 0.00146 nats and comparison RMSE changed by at most 0.00184 nats; correct-ordering probabilities changed by at most 2.5 percentage points for nonzero true differences. Thus global permutation centring left the principal KSG comparison errors unresolved in these settings. These descriptive results do not establish formal equivalence or apply to every randomisation reference (Online Resource 1, Section S6).

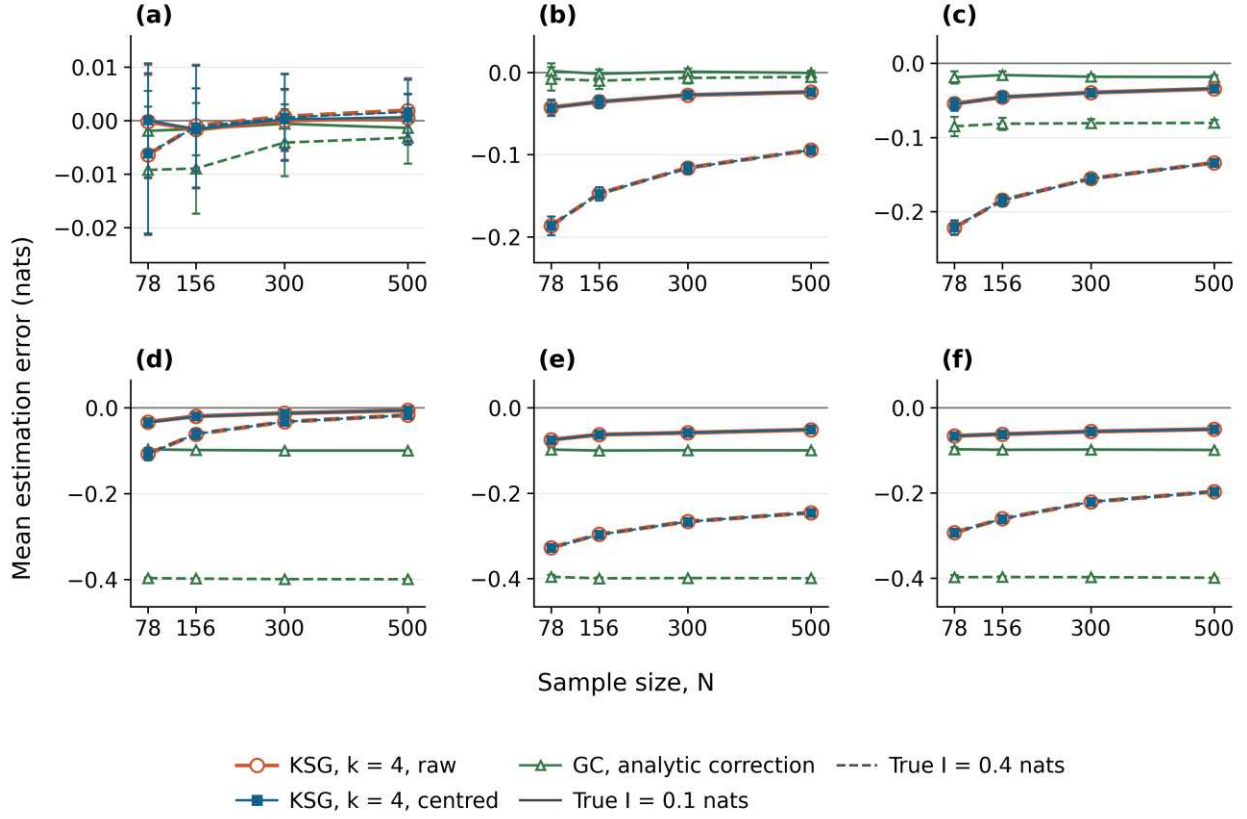


Fig. 1 Absolute calibration at matched true information. Mean error for raw and centred KSG with $k=4$ and analytically corrected Gaussian-copula scores. Panels (a)–(f) show L1, L3, T3, Q1, Q1x and Q3. Solid and dashed lines denote true information 0.10 and 0.40 nats. Hollow raw markers remain visible where curves overlap. Bars are 95% Student-t intervals across 200 independent parents; prefixes within each parent are nested. Vertical scales differ. The Gaussian correction is not generally exact after rank transformation

3.3 Between-sample comparisons

Equal-information samples in L3, T3, Q1, Q1x and Q3 had mean centred-KSG4 differences whose signs reversed when the sample-size imbalance was reversed (Fig. 2). Table 2 shows a nonzero-information comparison with $I^A=0.4$ and $I^B=0.6$ nats. At $N^A=500$ and $N^B=78$, L1 ordered the pair correctly in 92.5% of replicates, whereas Q1x did so in 13.0% (95% Wilson interval 9.0–18.4%). The contrast changes both dimension and geometry. More directly, Q1x had a mean difference of -0.0752 nats (95% interval -0.0841 to -0.0664), giving bias -0.2752 against the true $+0.2$. The squared empirical mean error was 95.0% of its empirical mean squared comparison error.

Common-78 matching raised Q1x correct ordering to 55.0% (48.1–61.7%) and reduced comparison RMSE from 0.2824 to 0.2015 nats. In the paired audit, 84 of 200 comparisons changed from incorrect to correct ordering and none changed in reverse; the gain was 42 percentage points (paired 95% percentile interval 35–49). Mean difference error nevertheless remained -0.1907 nats and the interval around 55.0% included 50%. Raw KSG4 gave 11.5% correct ordering before matching and 55.0% afterwards. At equal sizes, centred Q1x ordering was 56.0% for (78,78) and 86.0% for (500,500). Across designs, matched ordering lay between the equal-size (78,78) and (500,500) levels in seven of eight cases; Q1x was similar to the (78,78) level (55% versus 56%). Averaging five potentially overlapping subsets differs from using a single independent 78-row sample pair and retains parent-sample uncertainty. Matching provided a partial improvement that depended on absolute sample size and dependence structure.

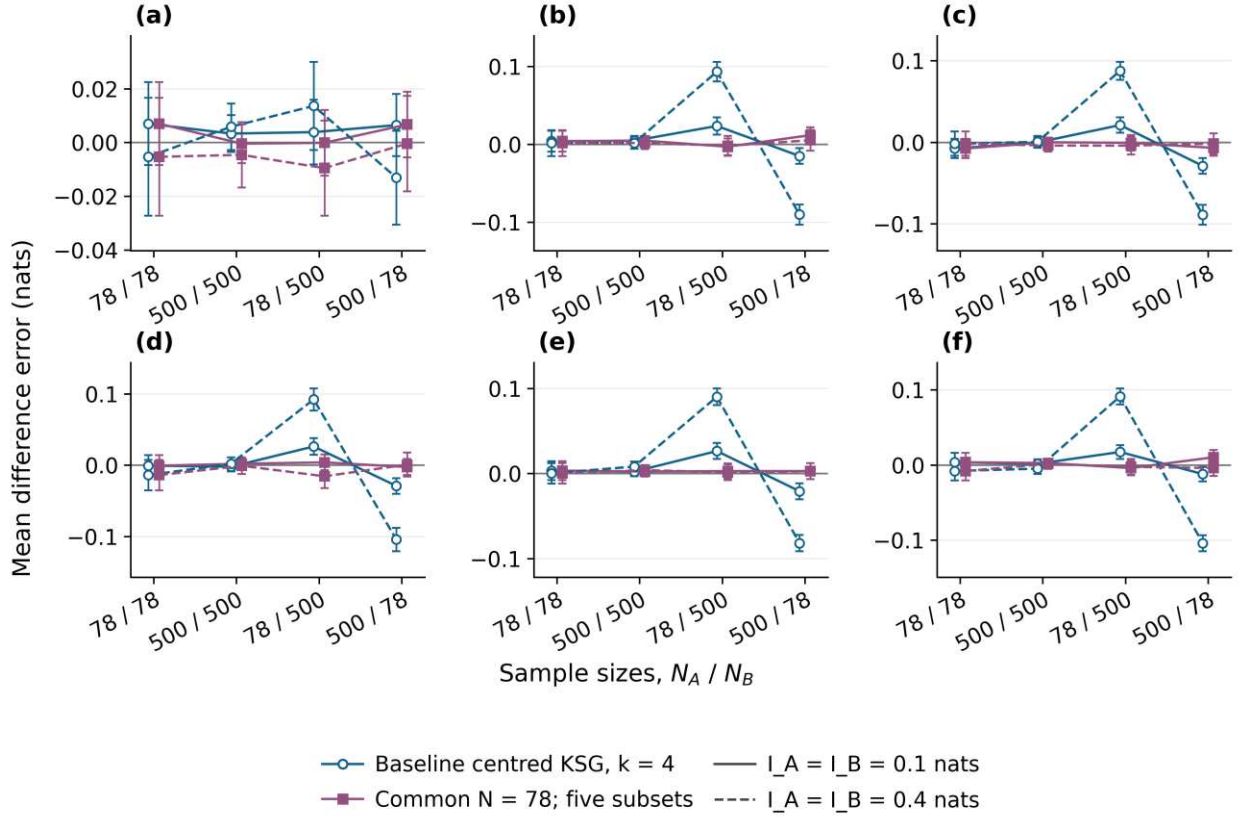


Fig. 2 Difference errors between independent samples with equal true information. Panels (a)–(f) show L1, L3, T3, Q1, Q1x and Q3 using centred KSG with $k=4$. The true difference is zero. Baseline estimates use the displayed sizes; matching uses 78 observations on both sides, averaging five subsets when the source has 500 rows. Solid and dashed lines denote information 0.10 and 0.40 nats. Bars are 95% Student-t intervals over 200 independent A/B parent pairs. Size comparisons reuse the same pairs and are not independent replications

Table 2 Comparison performance for true information $I^A=0.4$ and $I^B=0.6$ nats. Mean errors and root mean squared errors (RMSE) refer to $(N^A, N^B)=(500, 78)$, with baseline/common-78 values. Brackets give 95% Monte Carlo intervals for mean error. Correct-ordering probabilities at equal sizes use the original samples. Each row uses 200 independent parent pairs. H1 and H3 were reserved; the equal-size summaries are added in the revision audit. Full probability intervals are supplied in Online Resource 1, Section S6

Design	Mean error baseline	Mean error common-78	RMSE baseline/common	P(correct) baseline/common	P(correct) 78/78	P(correct) 500/500
L1	−0.030 [−0.0477, −0.0132]	−0.018 [−0.0352, −0.0003]	0.127 / 0.126	0.925 / 0.945	0.860	1.000
L3	−0.206 [−0.2190, −0.1928]	−0.110 [−0.1241, −0.0967]	0.226 / 0.148	0.530 / 0.820	0.800	0.995
T3	−0.186 [−0.1974, −0.1742]	−0.099 [−0.1104, −0.0872]	0.203 / 0.129	0.565 / 0.895	0.860	0.995
Q1	−0.180 [−0.1968, −0.1627]	−0.075 [−0.0928, −0.0567]	0.217 / 0.149	0.555 / 0.835	0.785	1.000
Q1x	−0.275 [−0.2841, −0.2664]	−0.191 [−0.1997, −0.1816]	0.282 / 0.201	0.130 / 0.550	0.560	0.860
Q3	−0.254 [−0.2654, −0.2420]	−0.154 [−0.1652, −0.1423]	0.267 / 0.174	0.240 / 0.695	0.665	0.950
H1	−0.093 [−0.1088, −0.0772]	−0.035 [−0.0511, −0.0184]	0.147 / 0.122	0.805 / 0.920	0.865	1.000
H3	−0.266 [−0.2757, −0.2559]	−0.154 [−0.1636, −0.1439]	0.275 / 0.169	0.190 / 0.750	0.695	0.985

3.4 Constant-information profiles and the subsampling tradeoff

At true information 0.4 nats, the mean centred-KSG4 profile rose by 0.0821 nats in Q1x as N increased from 78 to 500; its bootstrap endpoint interval was 0.0720–0.0922. Baseline profile ranges were outside the 0.04-nat tolerance in L3, T3, Q1, Q1x and Q3, and within it in L1 (Fig. 3). The common-78 profile was within tolerance in all six development designs at this target. At truth 0.1, all six common-78 profiles were indeterminate under the tighter 0.01-nat threshold.

The flatter profiles came with an absolute-calibration cost. For Q1 at $N=500$ and truth 0.4, common-78 changed the mean from 0.3826 to 0.2777 and RMSE from 0.0494 to 0.1339 nats. Its mean error moved from within to outside tolerance. Thus removing sample-size drift and recovering the true information magnitude were distinct outcomes. Common-size comparison results require their own error and RMSE assessment.

3.5 Equal-information geometry comparisons

Equal information and equal dimensions did not remove geometry effects (Fig. 4). At $N^A=N^B=500$ and $I=0.4$, the L3-to-T3 comparison had mean error -0.0394 nats (-0.0462 to -0.0326), and Q1x-to-Q3 had error $+0.0434$ (0.0367 to 0.0501). Common-78 values were -0.0354 and $+0.0344$. The intervals for these four comparisons overlapped the tolerance boundary, giving indeterminate classifications; this does not imply zero mean error.

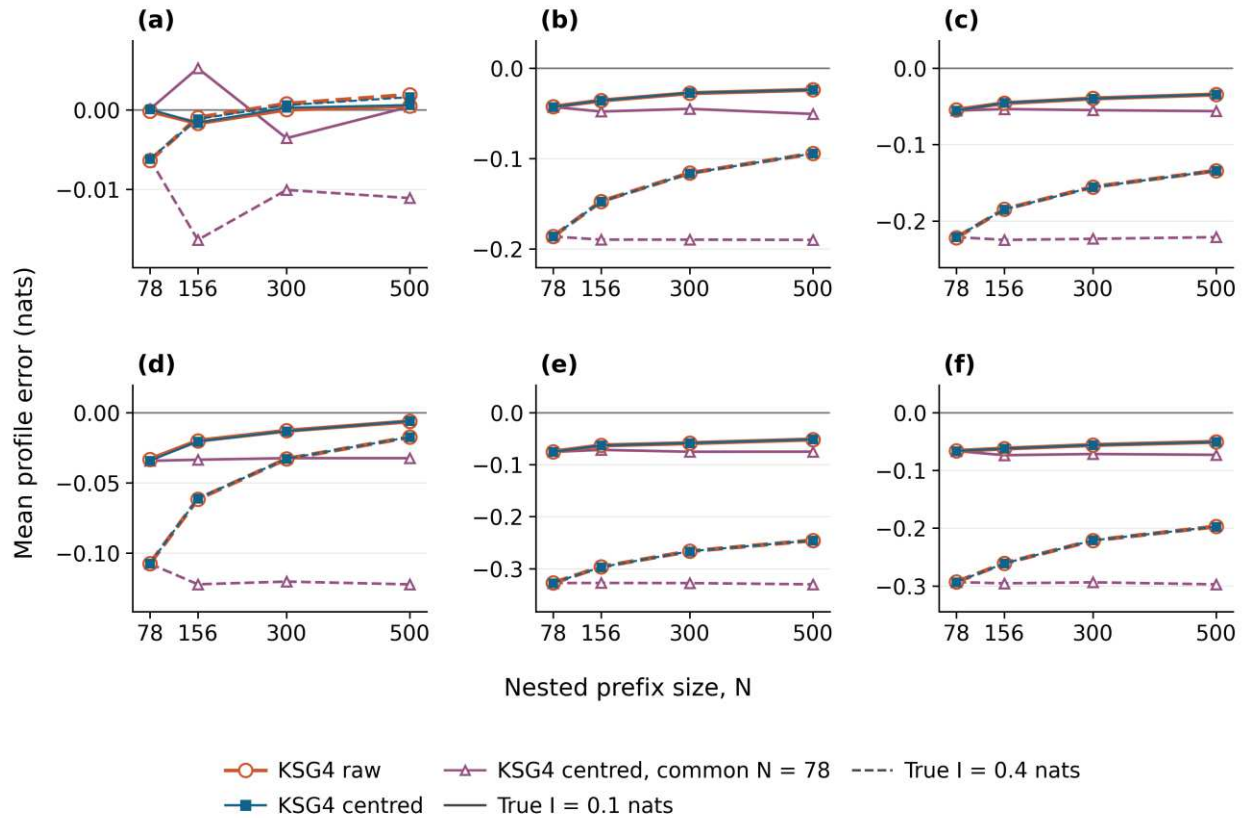


Fig. 3 Mean profiles under constant population information. Panels (a)–(f) show L1, L3, T3, Q1, Q1x and Q3. Curves give mean errors for raw and centred KSG with $k=4$ and centred common-78 matching. Solid and dashed lines denote true information 0.10 and 0.40 nats. Each curve uses 200 independent parents with nested prefixes; matching averages five subsets above $N=78$. Whole-parent bootstrap intervals for profile contrasts appear in the supplementary results. Vertical scales differ. A flat mean profile can remain biased in absolute magnitude

3.6 Reserved geometries and permutation sensitivity

The reserved absolute-value designs reproduced the comparison–calibration tradeoff (Table 2; Figs. S8–S11). In H3 at $I^A=0.4$, $I^B=0.6$, $N^A=500$ and $N^B=78$, the mean estimated difference was -0.0658 nats, despite the true $+0.2$. Common-78 reduced the mean difference error from -0.2658 to -0.1537 and RMSE from 0.2751 to 0.1692; correct ordering rose from 19.0% to 75.0%. At $I=0.4$, both reserved baseline profile ranges were outside tolerance and both

common-78 ranges were within it. At $I=0.1$, both common-78 profiles were indeterminate. In H1 at $N=500$ and $I=0.4$, subsampling increased absolute-estimation RMSE from 0.0480 to 0.0730. These are checks of fixed procedures on a reserved geometry, rather than validation of a fitted correction.

Across all 134 baseline design–target–stream–size settings, increasing the permutation count from 100 to 500 in the 20 preselected catalogues changed the mean centred-KSG4 estimate by at most 0.0029 nats in absolute value. The largest mean absolute paired change was 0.0055 nats. This sensitivity check concerns permutation Monte Carlo variation, not sampling bias, and does not extend the common-size calculations to 500 permutations. The complete $k=3,4,5,8$ analyses are supplied in the reproducibility tables and Figures S7 and S11; the primary k was not selected from their outcomes.

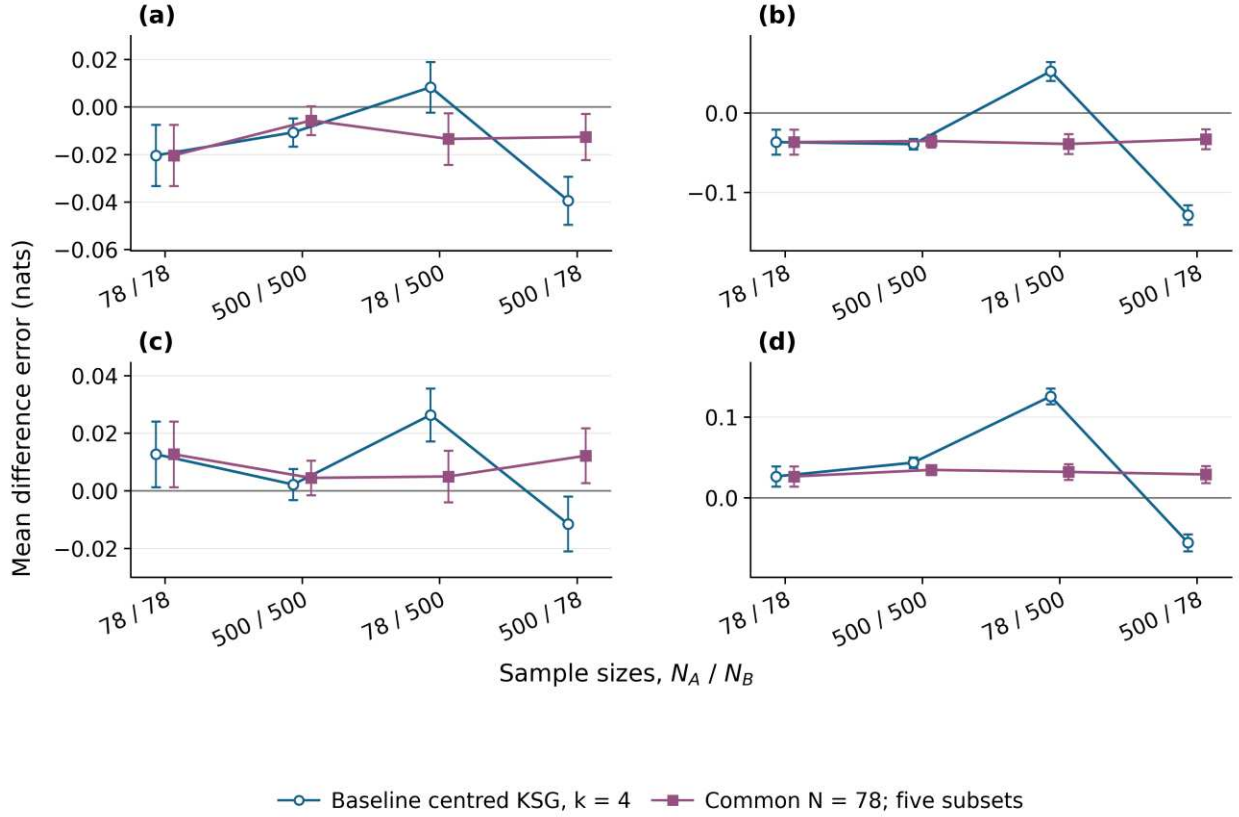


Fig. 4 Cross-geometry comparisons at equal information and dimensions (4,3), using centred KSG with $k=4$. Panels (a) and (b) compare $L3^A$ with $T3^B$ at true information 0.10 and 0.40 nats; panels (c) and (d) compare $Q1x^A$ with $Q3^B$ at those targets. $T3$ is an invertible deformation of $L3$; $Q1x$ and $Q3$ allocate information across one and three active pairs. Baseline and common-78 matching are defined as in Fig. 2. Bars are 95% Student-t intervals across 200 independent parent pairs. Equal information need not imply equal estimator bias

3.7 Moderate imbalance and minimum-size matching

The extension shows benefits and harms at moderate imbalances (Fig. 5). Across its complete grid of 64 unequal-size, nonzero-difference centred-KSG4 comparisons, five-subset minimum-size matching increased correct-ordering point estimates in 27 cells, decreased them in 34 and left three unchanged. Changes ranged from -33.6 to $+40.6$ percentage points. Pointwise paired 95% intervals excluded zero in 18 improving and 19 deteriorating cells. Comparison RMSE decreased in 34 cells and increased in 30; larger-catalogue absolute RMSE increased in 51. These are dependent grid summaries, not 64 independent studies or simultaneous significance statements.

A controlled comparison fixes $Q1x$ geometry, $N^A=200$, $N^B=400$ and an absolute information gap of 0.1 nats (Table 3). With $I^A=0.5$ and $I^B=0.4$, correct ordering rose from 40.8% to 63.9%, a paired increase of 23.1 percentage points (95% interval 19.8–26.5); comparison RMSE fell from 0.1301 to 0.0970 nats. Swapping only the information values gave 82.7% versus 60.7% correct ordering, a paired change of -22.0 points (-25.5 to -18.6), while comparison RMSE rose from 0.0742 to 0.1015 nats. The corresponding larger-sample

absolute RMSE rose from 0.2534 to 0.2876 nats and from 0.3320 to 0.3713 nats. Improvement in ordering therefore did not establish absolute calibration. For true differences -0.1 and $+0.1$ nats, the mean estimated differences changed from $+0.0171$ to -0.0181 and from $+0.0539$ to $+0.0139$ nats, respectively. Matching corrected the sign of the first mean but left substantial magnitude error in both directions.

Within Q1x, all 16 unequal cells with the stronger dependence in A had higher ordering point estimates after matching, and all 16 with the stronger dependence in B had lower ones. L1 changes were small and mixed; in the same 200-versus-400 comparison, paired ordering intervals included zero in both information directions. The directional Q1x pattern is conditional on this generator, estimator, sizes and five-subset procedure. It does not supply a general matching rule for unknown populations.

Permutation centring remained nearly neutral in the extension: mean KSG corrections ranged from -0.0009 to $+0.0005$ nats, the largest absolute raw-centred ordering change was 2.0 percentage points and the largest comparison-RMSE change was 0.00094 nats. All seven information pairs, both anchors, equal-size and equal-information controls, score variants, paired intervals and absolute errors are retained in Online Resource 2. Equal-information controls are interpreted through directional preference and difference error, without assigning a correct ordering.

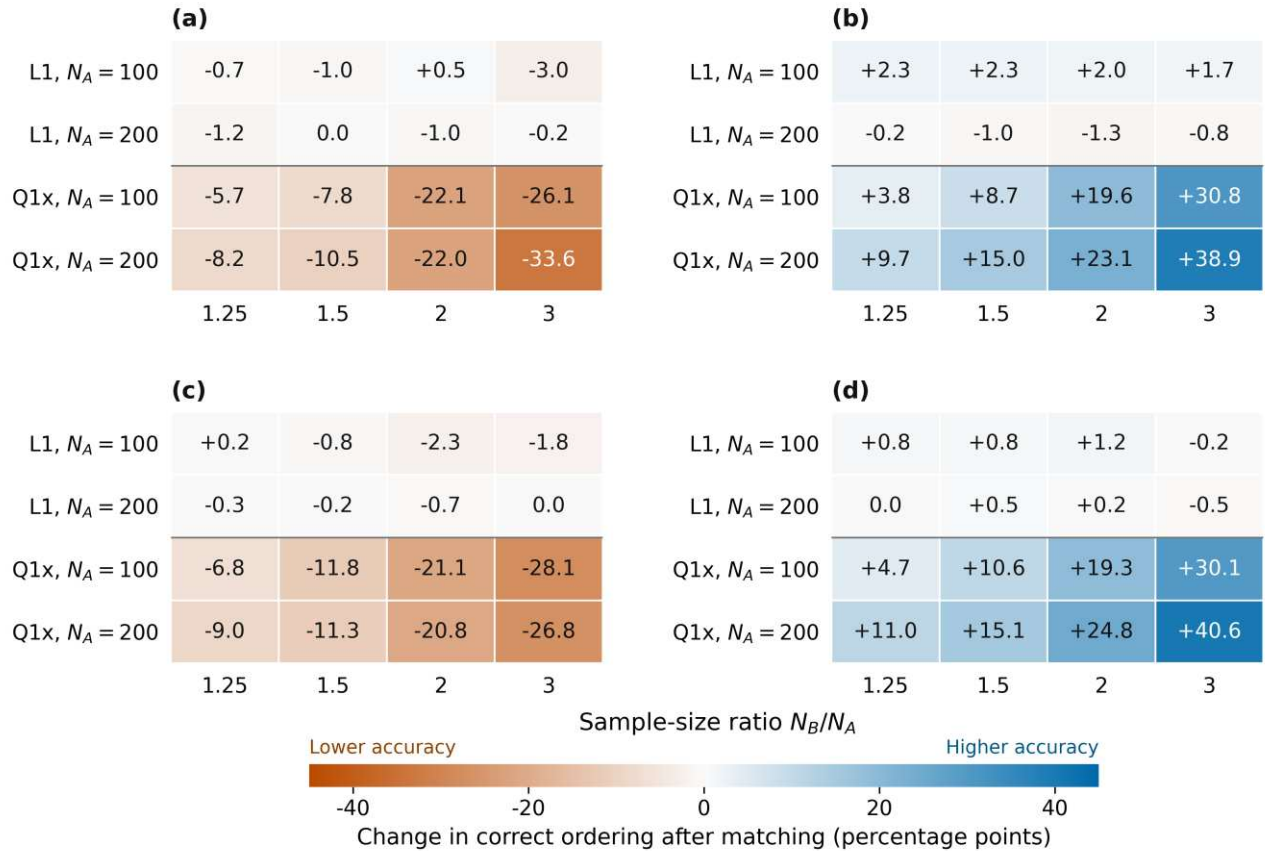


Fig. 5 Change in correct ordering under five-subset minimum-size matching. All 64 unequal-size, nonzero-difference extension cells are shown for centred KSG4. Panels (a)–(d) use true information pairs $(I^A, I^B) = (0.4, 0.5)$, $(0.5, 0.4)$, $(0.4, 0.6)$ and $(0.6, 0.4)$ nats, respectively. Rows identify geometry and the smaller size N^A ; N^B is the plotted ratio times N^A . Matching retains A and averages estimates from five size- N^A subsets of B, sampled without replacement within each subset; subsets may overlap. Each cell gives the matching-minus-baseline change in correct ordering, in percentage points, across 601 independent parent pairs. Positive values indicate improvement. Parent catalogues are reused across anchors, ratios and truth comparisons, so cells are dependent. Values are point estimates without significance symbols; their paired pointwise 95% bootstrap intervals do not all exclude zero and have no simultaneous-coverage guarantee. The plotted-value CSV and complete extension tables in Online Resource 2 supply ordering probabilities, intervals, comparison RMSE and absolute RMSE

Table 3 Controlled information swaps at sample sizes $N^A=200$ and $N^B=400$. Centred KSG4 is evaluated over 601 independent parent pairs per cell. Correct-ordering probabilities are percentages with pointwise Wilson 95% intervals; paired changes are matching minus baseline, in percentage points, with pointwise 95% parent-pair percentile intervals. RMSE is comparison root mean squared error in nats. Matching averages five B-subset estimates at N^A . Rows reuse banks; full-grid results are in Fig. 5 and Online Resource 2

Design	I^A / I^B (nats)	Baseline correct (%)	Matched correct (%)	Paired change (points)	Comparison RMSE baseline / matched
L1	0.4 / 0.5	87.9 [85.0, 90.2]	86.9 [83.9, 89.3]	-1.0 [-2.7, +0.7]	0.0832 / 0.0821
L1	0.5 / 0.4	89.7 [87.0, 91.9]	88.4 [85.5, 90.7]	-1.3 [-3.3, +0.5]	0.0881 / 0.0892
Q1x	0.4 / 0.5	82.7 [79.5, 85.5]	60.7 [56.8, 64.6]	-22.0 [-25.5, -18.6]	0.0742 / 0.1015
Q1x	0.5 / 0.4	40.8 [36.9, 44.7]	63.9 [60.0, 67.6]	+23.1 [+19.8, +26.5]	0.1301 / 0.0970

4 Stellar catalogue illustration

The archival illustration uses the VizieR catalogue access tool [25] and 94 distinct stars from CDS/VizieR catalogue J/A+A/511/L10, resolving duplicate observations by preferring Ultraviolet and Visual Echelle Spectrograph (UVES) measurements [12, 13]. A cut on the published local-standard-of-rest speed above 180 km s^{-1} selects 78 stars. The fixed chemical block is [Fe/H], [Mg/Fe], [Si/Fe] and [Ni/Fe]; the kinematic block is U, V and W. Selection, observations and features follow the archived observational study (Online Resource 1, Section S4). This illustration does not recalibrate an unknown stellar population using the synthetic designs.

The archived example distinguishes magnitude from evidence against a specified permutation reference (Table 4). KSG4, aligned with the simulation estimator, gives 0.2483 nats before centring. Global and instrument-by-metallicity-tertile references yield centred values 0.2381 and 0.1575 nats, with permutation p values 0.0020 and 0.0045. KSG5 and other statistics are retained as sensitivities. Restricted permutations preserve coarse metallicity structure and do not estimate conditional mutual information. These observations have unknown population information; they illustrate reference dependence without validating an unequal-sample correction.

The full observational selection, alternative nulls, neighbour orders, rounding and noise sensitivities are retained in Online Resource 1, Sections S4 and S6.4. Distance correlation is dimensionless [26], and the p values use the additional-count correction [27]. The example establishes neither a corrected latent information magnitude nor a new astrophysical population.

Table 4 Observed statistics and permutation-reference means for the 78-star illustration. Gaussian-copula (GC) and KSG scores are in nats; distance correlation is dimensionless. KSG4 is reconstructed from the archived streams and KSG5 retains the earlier sensitivity. The displayed p values are unadjusted, descriptive results for the specified nulls

Statistic	Observed	Global mean	Global p	Stratified mean	Stratified p
GC	0.1483	0.0803	0.0320	0.1601	0.5932
KSG4	0.2483	0.0103	0.0020	0.0909	0.0045
KSG5	0.2106	0.0080	0.0020	0.0838	0.0060
Distance correlation	0.4505	0.2914	0.0010	0.3492	0.0010

Global and instrument-by-metallicity-tertile references use 1,000 and 2,000 permutations, respectively. The $k=4$ stratified p value adjusted across the archived six neighbour orders by Holm’s method [28] is 0.0090. These references test different null hypotheses. Full distributions, centred values and reconstruction checks are supplied in Online Resource 1, Sections S4 and S6.4, and Online Resource 2.

5 Discussion

The effects of sample-size matching varied across the tested settings. In the Q1x extension at sample sizes 200 and 400, swapping the same two information values between the samples showed opposite effects: matching improved ordering when the stronger dependence was in the smaller sample and worsened it when the stronger dependence was in the larger sample. The full grid contains both benefits and harms, and matching often worsened absolute calibration for the larger sample. These results apply to the specified procedure and do not establish that matching is generally preferable or that a particular real-data ordering is trustworthy. Agreement with a full-data ranking, as assessed by Gao et al. [7], measures stability relative to that reference. Our known-truth comparisons assess correct ordering, difference error and absolute error separately.

Permutation centring measures the observed statistic relative to its randomisation ensemble. Under independent sampling and independence, exchangeability gives the centred statistic an expectation of zero. Under dependence, residual bias can depend on the estimator, sample size and joint distribution. A comparison then contains the difference between two residual biases. Reporting both estimates in nats does not make those biases equal.

A uniformly selected subset of an iid parent has the marginal distribution of an ordinary sample at the chosen size. Populations with identical complete joint distributions therefore have equal expected common-size estimates; equal mutual information alone does not imply this equality. Matching equates sample sizes while finite-sample biases can still differ. In the extension, A remains unchanged, so the shift in the mean estimated B-minus-A difference comes from the change in B's estimation bias. In the controlled Q1x swap, this downward shift corrects the mean ordering in one direction but moves the mean estimated difference further from truth in the other. Variance can also change. Averaging overlapping subsets reduces conditional subsampling variation but retains parent-sample uncertainty. The evaluated procedure combines this averaging with equalisation.

The GC results reflect a separate limitation. For the symmetric quadratic and absolute-value channels, population normal-score cross-covariances vanish even though mutual information is positive. The covariance score therefore omits this dependence at the population level. Neither an analytic finite-sample correction nor subtraction of a permutation mean can recover it. This symmetry does not imply a finite-sample rejection probability of exactly 5%; the independence-test simulations in Online Resource 1, Sections S3.1–S3.3, remain empirical observations for their specified generators.

Estimates from cumulative samples share observations, so a trend in a single realised profile need not indicate systematic estimator drift. We average profiles over independent parent catalogues and resample each complete prefix vector. This separates the range of the mean profile from its absolute displacement from truth. A flat profile may still be badly calibrated, while an individual noisy profile can have a large descriptive slope even when its mean is flat.

The centred benchmarks cover iid samples at $N=78$ –600, two block dimensions and the stated continuous mechanisms and information targets. The reserved absolute-value families in the original benchmark test fixed procedures on an additional geometry. The extension uses independently seeded catalogues but retains the previously studied L1 and Q1x geometries, with anchors 100 and 200, ratios at most three and only two nonzero information gaps. It broadens the sample-size comparison without establishing generality to arbitrary distributions. We did not evaluate serial dependence, unknown mixtures, measurement error or catalogue-selection corrections. The separate raw-score experiments at $N \leq 2000$ in Online Resource 1, Section S3, do not extend the centred benchmark.

The intervals describe Monte Carlo uncertainty in performance against known truth. They do not quantify uncertainty in an unknown population information difference for a single observed pair. Extension intervals and tests are pointwise, with no multiplicity control across the full grid. The fixed five-subset procedures combine equalisation and averaging, and the original 500-permutation sensitivity covers baseline estimates only. We did not implement a data-driven KSG diagnostic or parameter-selection comparator, derive a new correction, or validate a rule for selecting a procedure from real data. Ordinary observation bootstrap resampling can create duplicates that affect nearest-neighbour estimates [8]; our uncertainty calculations instead resample independent catalogue-level vectors.

In applications, report sample sizes, dimensions, estimator settings, the intended signed contrast and the randomisation reference. Where a defensible generative model is available, compare original-size estimates with explicitly defined matched procedures in a sensitivity analysis. Examine comparison RMSE, ordering and absolute errors together, and report adverse as well as favourable cases. The benchmark does not validate a choice between procedures based on the observed pair alone. Because the true information ordering and geometry are unknown in applications, the performance map is limited to the tested settings. A permutation p value addresses its specified null; calibrated magnitudes and reliable between-sample comparisons require separate evidence.

6 Conclusions

For the tested estimators, unequal sample sizes can reverse an average information ordering. In the two-anchor extension, five-subset minimum-size matching improved correct-ordering point estimates in 27 of 64 unequal comparisons and worsened them in 34. Swapping the stronger dependence between fixed sample sizes produced substantial effects in both directions. Comparison error and absolute calibration did not improve together, and global permutation centring left the principal KSG bias unresolved. The study quantifies the benefits and costs of the specified matching procedure, including adverse results and uncertainty. It does not establish a universal correction or a rule for judging reliability from a real-data pair.

Statements and Declarations

Funding. No external funding was received for this work.

Competing interests. The authors declare no competing interests.

Ethics approval and consent. Not applicable. This study uses synthetic data and publicly available stellar catalogue measurements.

Author contributions. Nick Barua: Conceptualisation, Methodology, Software, Formal analysis, Visualisation, Writing – original draft, Writing – review and editing. Koji Oda: Validation, Writing – review and editing.

Data and code availability. The reproducibility archive (Online Resource 2; ESM_2.zip) is available on Zenodo at <https://doi.org/10.5281/zenodo.23466586> [29]. It contains frozen protocols, code, seed specifications, procedure-level replicate estimates, extension comparison vectors, truth checks, audits, performance tables, stellar-analysis materials, plotting code and checksums. Individual estimator-call checkpoints and redundant exports are omitted; checkpoints can be regenerated with the supplied code and seeds. The source stellar measurements are in CDS/VizieR catalogue J/A+A/511/L10 [13]. Online Resource 1 supplies supplementary methods, validation and extension reporting. No new observations were collected.

Online Resources

Online Resource 1 (ESM_1.pdf) Supplementary mathematical identities, simulation methods, numerical validation, original and reserved results, retrospective audits, stellar-catalogue sensitivities and the moderate-imbalance extension

Online Resource 2 Reproducibility archive (ESM_2.zip), deposited on Zenodo [29]

References

- [1] Paninski L (2003) Estimation of entropy and mutual information. *Neural Comput* 15:1191–1253. <https://doi.org/10.1162/089976603321780272>
- [2] Gao W, Oh S, Viswanath P (2018) Demystifying fixed k-nearest neighbor information estimators. *IEEE Trans Inf Theory* 64:5629–5661. <https://doi.org/10.1109/TIT.2018.2807481>
- [3] Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. *Phys Rev E* 69:066138. <https://doi.org/10.1103/PhysRevE.69.066138>
- [4] 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. *Hum Brain Mapp* 38:1541–1573. <https://doi.org/10.1002/hbm.23471>

- [5] Khan S, Bandyopadhyay S, Ganguly AR, Saigal S, Erickson DJ III, Protopopescu V, Ostrouchov G (2007) Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Phys Rev E* 76:026209. <https://doi.org/10.1103/PhysRevE.76.026209>
- [6] Czyż P, Grabowski F, Vogt J, Beerenwinkel N, Marx A (2023) Beyond Normal: On the Evaluation of Mutual Information Estimators. *Adv Neural Inf Process Syst* 36:16957–16990. <https://doi.org/10.52202/075280-0742>
- [7] Gao S, Ver Steeg G, Galstyan A (2015) Efficient estimation of mutual information for strongly dependent variables. In: Lebanon G, Vishwanathan SVN (eds) *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*. PMLR 38:277–286. <https://proceedings.mlr.press/v38/gao15.html>
- [8] Holmes CM, Nemenman I (2019) Estimation of mutual information for real-valued data with error bars and controlled bias. *Phys Rev E* 100:022404. <https://doi.org/10.1103/PhysRevE.100.022404>
- [9] Sheintuch L, Rubin A, Ziv Y (2022) Bias-free estimation of information content in temporally sparse neuronal activity. *PLoS Comput Biol* 18:e1009832. <https://doi.org/10.1371/journal.pcbi.1009832>
- [10] Barua N (2026) Held-out dynamical validation of a cross-survey chemical decomposition of 468,000 Milky Way stars. SSRN preprint. <https://doi.org/10.2139/ssrn.7385521>
- [11] Pandey B, Mondal A (2026) Stellar Halo Memory: A New Observable for Galactic Archaeology. arXiv preprint arXiv:2607.10863. <https://doi.org/10.48550/arXiv.2607.10863>
- [12] Nissen PE, Schuster WJ (2010) Two distinct halo populations in the solar neighborhood. Evidence from stellar abundance ratios and kinematics. *Astron Astrophys* 511:L10. <https://doi.org/10.1051/0004-6361/200913877>
- [13] Nissen PE, Schuster WJ (2010) Abundances and space velocities of 94 stars [data set]. CDS/VizieR catalogue J/A+A/511/L10. <https://doi.org/10.26093/cds/vizier.35119010>
- [14] Morris TP, White IR, Crowther MJ (2019) Using simulation studies to evaluate statistical methods. *Stat Med* 38:2074–2102. <https://doi.org/10.1002/sim.8086>
- [15] Cover TM, Thomas JA (2006) *Elements of information theory*, 2nd edn. Wiley, Hoboken. <https://doi.org/10.1002/047174882X>
- [16] Ernst MD (2004) Permutation methods: A basis for exact inference. *Stat Sci* 19(4):676–685. <https://doi.org/10.1214/088342304000000396>
- [17] Kozachenko LF, Leonenko NN (1987) Sample estimate of the entropy of a random vector. *Probl Inf Transm* 23(2):95–101. <https://www.mathnet.ru/eng/ppi797>
- [18] Wilson EB (1927) Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* 22:209–212. <https://doi.org/10.1080/01621459.1927.10502953>
- [19] Brown LD, Cai TT, DasGupta A (2001) Interval estimation for a binomial proportion. *Stat Sci* 16(2):101–133. <https://doi.org/10.1214/ss/1009213286>
- [20] Koehler E, Brown E, Haneuse SJ-PA (2009) On the assessment of Monte Carlo error in simulation-based statistical analyses. *Am Stat* 63(2):155–162. <https://doi.org/10.1198/tast.2009.0030>
- [21] Harris CR, Millman KJ, van der Walt SJ et al (2020) Array programming with NumPy. *Nature* 585:357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- [22] Virtanen P, Gommers R, Oliphant TE et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 17:261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- [23] McKinney W (2010) Data structures for statistical computing in Python. In: van der Walt S, Millman J (eds) *Proceedings of the 9th Python in Science Conference*, pp 56–61. <https://doi.org/10.25080/Majora-92bf1922-00a>
- [24] Hunter JD (2007) Matplotlib: a 2D graphics environment. *Comput Sci Eng* 9(3):90–95. <https://doi.org/10.1109/MCSE.2007.55>
- [25] Ochsenbein F, Bauer P, Marcout J (2000) The VizieR database of astronomical catalogues. *Astron Astrophys Suppl Ser* 143:23–32. <https://doi.org/10.1051/aas:2000169>
- [26] Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. *Ann Stat* 35:2769–2794. <https://doi.org/10.1214/009053607000000505>
- [27] Phipson B, Smyth GK (2010) Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol* 9:Article 39. <https://doi.org/10.2202/1544-6115.1585>
- [28] Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2):65–70. <https://www.jstor.org/stable/4615733>
- [29] 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>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESM1.pdf](#)
- [ESM2.zip](#)