DIFFERENTIABLE ESTIMATION OF CONDITIONAL MUTUAL INFORMATION FOR MIXED-TYPE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional mutual information (CMI) is a fundamental measure of statistical dependence, but using it as an optimization objective for mixed categorical and continuous data has been limited by the lack of estimators that are both consistent and differentiable. In this article, we devise a product-kernel CMI estimator that combines exact categorical matching with Gaussian kernel smoothing over continuous coordinates. We prove its consistency for a canonical mixed setting under standard bandwidth conditions and give an explicit error decomposition into smoothing bias, local-kernel stochastic error, and empirical-averaging error. We further introduce a differentiable relaxation that replaces hard categorical matching with soft probability-vector similarities compatible with gradient-based optimization. We show that, when the relaxed representations approach their hard one-hot counterparts at rate a_n, the soft estimator deviates from the hard one by $O_p(a_n)$; thus, consistency of the hard estimator carries over to its differentiable counterpart. Finally, we show consistency for both the hard and differentiable estimators across all eight possible categorical–continuous configurations of the three variables involved in a CMI query. Our results provide theoretical guarantees for using CMI as a differentiable regularizer in gradient-based procedures for mixed-type data, such as constrained data imputation, fairness-aware learning, and invariant representation learning.

1 INTRODUCTION

Controlling statistical dependence is often critical in learning and data-repair problems. For example, fairness constraints may require predictions to be independent of sensitive attributes conditional on outcomes (Hardt et al., 2016); data-repair methods may enforce conditional-independence constraints (Pirhadi et al., 2024); and information-based regularization can suppress unwanted dependence in learned representations (Nguyen et al., 2022). In these settings, conditional independence must be included in the optimization of learning procedures.

To express such requirements, we commonly use conditional mutual information (CMI), defined as

$$I(U;V \mid W) = \mathbb{E}\left[\log \frac{p(U,V \mid W)}{p(U \mid W)p(V \mid W)}\right]$$

for random variables U, V, and W, where $p(\cdot \mid \cdot)$ denotes the probability mass or density functions for the random variables indicated by the corresponding arguments. CMI has an appealing property in reflecting conditional independence by assuming the value zero. That is, under standard regularity conditions, $I(U;V \mid W) = 0 \iff U \perp V \mid W$. CMI is thereby often used as a penalty in objectives of the form

$$\min_{\theta} \left[\mathcal{L}_{\text{task}}(\theta) + \lambda I(U;V \mid W;\theta)\right], \tag{1}$$

where $\mathcal{L}_{\text{task}}(\theta)$ is the loss function constructed to facilitate a particular learning task, indexed by parameter θ to be optimized, λ is a nonnegative tuning parameter, and U, V, and W are the associated random variables in the learning task.

While (1) provides a convenient principled framework, its implementation requires estimation of CMI. With categorical variables only, we can obtain a closed-form expression of CMI, whereas for

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

continuous variables involved, we may use local smoothing to approximate CMI. However, when data contain mixed-type of categorical and continuous variables, estimating CMI is not straightforward. Existing estimators for mixed data use nearest-neighbor, histogram, or hybrid constructions (Mesner & Shalizi, 2021; Marx et al., 2021; Zan et al., 2022). These methods address CMI estimation or conditional-independence testing from observed data. Optimization adds an additional difficulty: if a categorical representation depends on θ, an exact-match indicator is discontinuous and provides no gradient.

To address these challenges, we develop a product-kernel estimator, Product-CMI. In its hard form, categorical coordinates contribute exact-match factors and continuous coordinates contribute Gaussian similarities. The resulting products estimate the four local quantities in the CMI ratio. For optimization, we replace each categorical match by the inner product of two categorical probability vectors. This recovers exact matching for one-hot vectors and remains differentiable for soft representations. We further describe statistical links between the hard and soft estimators, and prove consistency of the hard estimator for all eight categorical–continuous assignments of (U, V, W). In a representative mixed setting, we decompose the estimation error into smoothing bias, local stochastic error, and empirical-averaging error. If the soft categorical representations approach their one-hot counterparts at a rate, say a_n, the differentiable estimator differs from the hard estimator by $O_p(a_n)$. The relaxation therefore preserves the limiting target without introducing an additional inverse-bandwidth factor. Our main contributions are summarized as follows:

- **Mixed-type Product-CMI:** We introduce a CMI estimator based on exact categorical matching and Gaussian smoothing over continuous coordinates. We prove consistency across all eight type assignments of (U, V, W) and derive an explicit first-order error decomposition for a representative mixed setting.
- **A differentiable categorical relaxation:** We replace exact categorical matches by similarities between probability vectors and show that $\widehat{I}_n^{\mathrm{S}} - \widehat{I}_n^{\mathrm{H}} = O_p(a_n)$ when the soft representations approach their one-hot counterparts at rate a_n, where $\widehat{I}_n^{\mathrm{S}}$ and $\widehat{I}_n^{\mathrm{H}}$ represent the soft and hard Product-CMI estimators, respectively.
- **Evaluation in estimation and optimization:** We compare Product-CMI with mixed-data CMI baselines, evaluate the finite-sample performance, and test whether the differentiable objective moves parameters toward known conditional-independence solutions.

2 Related Work

CMI estimation for mixed data. Classical nonparametric estimators of mutual information and CMI are often based on nearest neighbors or density estimation, and they are usually developed for handling continuous variables (Kraskov et al., 2004; Frenzel & Pompe, 2007). Several methods extend CMI estimation to data containing both categorical and continuous components. For example, Mesner & Shalizi (2021) develop a nearest-neighbor estimator for mixed distributions and establish its consistency. Their construction is one of our main statistical baselines because it directly targets mixed-type CMI without first discretizing all continuous variables. Marx et al. (2021) take a different approach based on multidimensional adaptive histograms, where discretization of the continuous coordinates is learned from the data. Zan et al. (2022) propose CMIh, which combines histogram-based estimation for qualitative variables with nearest-neighbor estimation for quantitative variables and is designed for both CMI estimation and conditional-independence testing. Some recent work continues to explore nearest-neighbor methods for mixed-variable conditional-independence testing (Popescu et al., 2025).

These methods are closely related to our proposed Product-CMI in that they estimate CMI from mixed categorical–continuous samples. For comparisons, in our subsequent experiments, we include both Mesner–Shalizi and CMIh as baseline methods, and we also evaluate the official Marx implementation on the subset of benchmark configurations that it supports. However, our method differs from existing works whose primary objective is estimation or conditional-independence testing on an observed dataset. In contrast, our Product-CMI is designed with an additional requirement: the same construction should be usable when categorical representations are optimization variables. Its hard form uses exact categorical matching, while its soft form replaces those matches by differentiable

similarities between probability vectors. Our analysis then relates the soft estimator back to the hard mixed-type estimator.

Kernel methods for heterogeneous data. The use of product kernels for mixtures of categorical and continuous variables is well established in nonparametric statistics. To name a few, Li & Racine (2003), Hall et al. (2004), and Racine & Li (2004) develop kernel methods for mixed-data density, conditional-density, and regression estimation using type-specific kernels. Our proposed Product-CMI builds on this general principle. While it is standard to use Gaussian smoothing or separate categorical and continuous kernel factors, the novelty of our method lies in how these local mixed-type similarities are effectively assembled into the four quantities required by the CMI ratio. We rigorously establish consistency across all possible eight combinations of variable nature for (U, V, W) and uncover the statistical connection between the exact-match estimator and its differentiable relaxation.

Differentiable information estimation and categorical relaxations. A separate line of work develops differentiable estimators and bounds for information-theoretic quantities. MINE uses a neural variational estimator of mutual information (Belghazi et al., 2018), and later work studies variational bounds and neural estimators for mutual information and CMI (Poole et al., 2019; Mukherjee et al., 2020; Mondal et al., 2020; Molavipour et al., 2020). DINE provides another neural information estimator based on diffeomorphic transformations and is formulated for continuous variables (Duong & Nguyen, 2023). Gumbel–Softmax and the Concrete distribution provide standard continuous relaxations for categorical variables (Jang et al., 2017; Maddison et al., 2017).

Our setting sits between these two lines of work. We require an estimator that retains explicit mixed categorical–continuous semantics and also admits gradients when the categorical representations are relaxed. To the best of our knowledge, our work is the first one that extends prior works, which do not combine mixed-type CMI estimation, a differentiable relaxation of the categorical comparisons, and an explicit statistical guarantee relating the relaxed estimator to its hard mixed-type counterpart.

3 Product-CMI: Differentiable Mixed-Type Estimation

3.1 Hard Mixed-Type Estimator

Let (U_i, V_i, W_i), $i = 1, \ldots, n$, be independent and identically distributed (i.i.d.) random variables following the same distribution as the random vector (U, V, W). Each of U, V, and W may be categorical or continuous, where categorical variables have finite support and continuous variables may be vector-valued. For $X \in \{U, V, W\}$ and $i, j = 1, \ldots, n$, define the pairwise factor for X_i and X_j

$$\kappa_{ij}^{X}(h) = \begin{cases} \mathbf{1}\{X_i = X_j\}, & \text{if } X \text{ categorical,} \\ \exp\left(-\dfrac{\|X_i - X_j\|_2^2}{2h^2}\right), & \text{if } X \text{ continuous.} \end{cases} \tag{2}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, and h denotes the bandwidth parameter. This formulation reflects that categorical variables contribute through exact matching, and continuous variables contribute through Gaussian similarity. For any nonempty $S \subseteq \{U, V, W\}$ and $i, j = 1, \ldots, n$, define

$$K_{S,ij}^{\mathrm{H}}(h) = \prod_{X \in S} \kappa_{ij}^{X}(h) \qquad \text{and} \qquad M_{S,i}^{\mathrm{H}}(h) = \frac{1}{n-1} \sum_{j \neq i} K_{S,ij}^{\mathrm{H}}(h),$$

where the superscript H denotes the hard estimator. The quantity $M_{S,i}^{\mathrm{H}}$ is a local mixed-type mass around observation i.

Then using the four local masses required by the CMI ratio, we define

$$\widehat{r}_i^{\mathrm{H}} = \frac{M_{UVW,i}^{\mathrm{H}} M_{W,i}^{\mathrm{H}}}{M_{UW,i}^{\mathrm{H}} M_{VW,i}^{\mathrm{H}}} \qquad \text{and} \qquad \widehat{I}_n^{\mathrm{H}} = \frac{1}{n} \sum_{i=1}^{n} \log \widehat{r}_i^{\mathrm{H}}.$$

Observation i is excluded from its own local masses so that deterministic self-similarity does not contribute to the estimate. A single bandwidth h is shared by all continuous Gaussian factors; this common scale is also used in the consistency argument of Section 4.

3.2 Differentiable Categorical Relaxation

Exact categorical matching is discontinuous when the categorical representation itself depends on parameters being optimized. Suppose a categorical variable X has K_X levels and observation i is represented by $p_i^X \in \Delta^{K_X-1}$, where Δ^{K_X-1} is the probability simplex. We replace the hard categorical match by

$$s_{ij}^X = (p_i^X)^\top p_j^X. \tag{3}$$

For one-hot vectors, (3) is exactly the hard match indicator. For soft probability vectors, it varies smoothly with the representation.

The soft product factors are obtained by replacing each categorical factor in (2) with (3), while leaving the continuous Gaussian factors unchanged. Let $M_{S,i}^{\mathrm{S}}$ denote the resulting local masses, and then in contrast to the form of $\widehat{I}_n^{\mathrm{H}}$, we define

$$\widehat{I}_n^{\mathrm{S}} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{M_{UVW,i}^{\mathrm{S}} M_{W,i}^{\mathrm{S}}}{M_{UW,i}^{\mathrm{S}} M_{VW,i}^{\mathrm{S}}}, \tag{4}$$

where the superscript S indicates this is a soft estimator.

For one-hot categorical representations, this recovers the hard estimator exactly, i.e., $\widehat{I}_n^{\mathrm{S}} = \widehat{I}_n^{\mathrm{H}}$. For probability vectors in the relative interiors of their simplices, the soft categorical similarities are positive and differentiable. Together with the Gaussian factors, this makes (4) differentiable with respect to the relaxed categorical representations and continuous coordinates used as optimization variables. In Section 4, we further quantify the difference between the hard and soft estimators away from the one-hot limit.

3.3 Practical Implementation

In our subsequent experiments, we use a shared scalar bandwidth for the continuous Gaussian factors. Configuration-specific preprocessing and the exact bandwidth-search procedure are given in Appendix B; reference CMI is never used for bandwidth selection.

In sparse finite samples, exact categorical matching can produce very small local masses. Before taking logarithms, we therefore floor each local mass at approximately 10^{-12}. The final CMI estimate is not clipped. The theoretical results in Section 4 concern the unfloored estimator. A direct implementation evaluates pairwise similarities and has $O(n^2)$ time complexity. The required sums can be accumulated blockwise when memory is limiting without changing the estimator. In optimization problems, we replace the population CMI in (1) with the differentiable estimator (4). The dependence on θ may enter through continuous values, categorical probability vectors, or both.

4 Statistical Guarantees

We establish the statistical properties for hard and soft Product-CMI, with the assumptions and main results included here and complete proofs deferred to Appendix A.

4.1 Assumptions and Local-Mass Approximation

For $\mathcal{S} = \{UVW, UW, VW, W\}$, let g_S denote the corresponding probability mass function, density, or mixed mass–density, with counting measure for categorical coordinates and Lebesgue measure for continuous ones. For $X \in \{U, V, W\}$, let d_X denote its Euclidean dimension when X is continuous and set $d_X = 0$ when X is categorical. Define $D = d_U + d_V + d_W$, $d_S = \sum_{X \in S} d_X$, and $b_S(h) = (2\pi)^{d_S/2} h^{d_S}$, with $b_S(h) = 1$ when $d_S = 0$. The population CMI ratio is then given by

$$r(u, v, w) = \frac{g_{UVW}(u, v, w) g_W(w)}{g_{UW}(u, w) g_{VW}(v, w)}, \qquad I(U; V \mid W) = \mathbb{E}[\log r(U, V, W)]. \tag{5}$$

The following regularity conditions are imposed.

A1 **(Sampling).** The observations (U_i, V_i, W_i), $i = 1, \ldots, n$, are i.i.d.

A2 **(Categorical support).** Every categorical coordinate has fixed finite support, and every categorical configuration relevant to the four marginals in $\mathcal{S}$ has positive probability.

A3 **(Smoothness and positivity).** On a fixed region $\mathcal{X}_0$, the relevant mixed marginals are bounded above and away from zero. Along continuous coordinates they are twice continuously differentiable with uniformly bounded second derivatives.

A4 **(Log-ratio integrability).** $\mathbb{E}[|\log r(U, V, W)|] < \infty$.

A5 **(Bandwidth).** To stress the dependence on the sample size n, we write h as h_n. When $D > 0$, assume that $h_n \to 0$ and $\frac{nh_n^D}{\log n} \to \infty$ as $n \to \infty$.

Condition A5 is not needed when all three variables are categorical. The condition $h_n \to 0$ removes smoothing bias, and the condition $\frac{nh_n^D}{\log n} \to \infty$ provides sufficient local sample support. The following lemma gives the common form of the four local Product masses.

Lemma 1 (Local-mass approximation)**.** *Suppose Assumptions A1–A5 hold. Then, for every $S \in \mathcal{S}$,*

$$M_{S,i}^{\mathrm{H}}(h_n) = b_S(h_n) g_S(X_{S,i})\{1 + o_p(1)\},$$

uniformly over sample points and simultaneously over the four choices of S.

Categorical coordinates restrict the local average to matching strata, while the continuous coordinates contribute a d_S-dimensional Gaussian localization.

4.2 Theoretical Results

The key identity $d_{UVW} + d_W = d_{UW} + d_{VW}$ implies

$$b_{UVW}(h_n) b_W(h_n) = b_{UW}(h_n) b_{VW}(h_n). \tag{6}$$

By Lemma 1, each local mass is its population counterpart times $b_S(h_n)\{1 + o_p(1)\}$. The b_S factors cancel by (6), giving $\widehat{r}_i^{\mathrm{H}} = r(U_i, V_i, W_i)\{1 + o_p(1)\}$. Positivity permits the logarithm, and the empirical average converges by the law of large numbers.

Theorem 1 (Hard Product-CMI consistency)**.** *Suppose Assumptions A1–A5 hold. Then*

$$\widehat{I}_n^{\mathrm{H}} \xrightarrow{p} I(U; V \mid W) \quad \textit{as} \quad n \to \infty.$$

for all eight combinations of categorical and continuous variables in (U, V, W).

For a more explicit error decomposition, consider categorical U and W with $V \in \mathbb{R}^d$. Let

$$\pi_{\min} = \min_{u,w:\, \mathbb{P}(U=u, W=w)>0} \mathbb{P}(U = u, W = w).$$

In addition to Assumptions A1–A5, we further assume

$$\mathbb{E}[(\log r(U, V, W))^2] < \infty \quad \text{and} \quad \frac{n\pi_{\min} h_n^d}{\log n} \to \infty \quad \text{as } n \to \infty.$$

Then

$$\widehat{I}_n^{\mathrm{H}} - I(U; V \mid W) = O_p\left(h_n^2 + \sqrt{\frac{\log n}{n\pi_{\min} h_n^d}} + n^{-1/2}\right). \tag{7}$$

The terms in the order O_p correspond respectively to smoothing bias, local stochastic error, and empirical averaging error over a sample of size n. The factor $n\pi_{\min} h_n^d$ acts as an effective local sample size: categorical matching contributes $\pi_{\min}$ and continuous localization contributes h_n^d. This explains why rare strata and higher continuous dimension make local estimation harder. Expression (7) is a first-order error decomposition, and it is not necessarily a sharp or minimax-optimal rate.

For a categorical variable X, let e_{X_i} denote the one-hot representation of the observed category and define

$$\delta_n = \max_{\substack{X \in \{U,V,W\} \\ X \text{ categorical}}} \max_{1 \le i \le n} \|p_i^X - e_{X_i}\|_1,$$

with $\delta_n = 0$ when all variables are continuous. The quantity δ_n measures the largest departure of any soft categorical representation from its observed one-hot value. To see how this departure affects the CMI estimator, we compare the local kernel masses computed with soft categorical similarities and with exact categorical matching. The soft perturbation has the same kernel scale as the corresponding hard local mass. Uniformly over i and $S \in \mathcal{S}$,

$$\left|M_{S,i}^{\mathrm{S}}(h_n) - M_{S,i}^{\mathrm{H}}(h_n)\right| = O_p\left(\delta_n b_S(h_n)\right), \qquad M_{S,i}^{\mathrm{H}}(h_n) \asymp b_S(h_n). \tag{8}$$

Hence the corresponding relative error is $O_p(\delta_n)$. The common kernel-volume scale cancels, so the relaxation does not introduce an additional inverse power of the bandwidth.

Theorem 2 (Soft-to-hard consistency transfer)**.** *Suppose that Assumptions A1–A5 hold and that, for some deterministic sequence $a_n \to 0$, $\delta_n = O_p(a_n)$. Then $\widehat{I}_n^{\mathrm{S}} - \widehat{I}_n^{\mathrm{H}} = O_p(a_n)$. Consequently, $\widehat{I}_n^{\mathrm{S}} \xrightarrow{p} I(U;V \mid W)$ as $n \to \infty$.*

Under the conditions for (7), combining Theorem 2 with (7) gives

$$\widehat{I}_n^{\mathrm{S}} - I(U;V \mid W) = O_p\left(h_n^2 + \sqrt{\frac{\log n}{n\pi_{\min} h_n^d}} + n^{-1/2} + a_n\right). \tag{9}$$

Comparing (9) to (7) shows that the difference of $\widehat{I}_n^{\mathrm{S}}$ and $\widehat{I}_n^{\mathrm{H}}$ from $I(U;V \mid W)$ is reflected by the a_n term, which is due to the categorical relaxation. The result in (9) is conditional on the soft representations approaching their one-hot counterparts; it does not imply convergence of a particular optimizer or temperature schedule.

5 Empirical Evaluation

We evaluate the performance of the proposed Product-CMI in terms of estimator accuracy, the finite-sample effects, and optimization toward a known conditional-independence solution. The synthetic benchmark uses six estimator type configurations: (C,C,C), (C,C,N), (C,N,C), (C,N,N), (N,N,C), and (N,N,N), where C denotes categorical treatment and N numerical treatment. They cover the eight assignments up to the symmetry $I(U;V \mid W) = I(V;U \mid W)$. We combine them with four dependence levels, binary and multiclass data-generating families, and five random seeds. We use $n \in \{500, 1000, 2000, 5000, 10000\}$, giving 240 runs per sample size. Reference CMI is computed independently and is not used for estimator tuning. Details about data-generating mechanisms, preprocessing, bandwidth selection, and reference-CMI calculations are given in Appendix B.

5.1 Estimator Accuracy Across Sample Sizes

We compare hard Product-CMI with Mesner–Shalizi (Mesner & Shalizi, 2021) and CMIh (Zan et al., 2022) on the same 240 runs at each sample size. Mesner–Shalizi uses the Tigramite implementation with fixed $k = 7$, and CMIh uses the official `mixedEstimator` implementation. Accuracy is measured by mean absolute error (MAE) from the independently computed reference CMI, in nats.

Figure 1 shows a strong sample-size effect for Product-CMI: pooled MAE decreases from 0.1054 at $n = 500$ to 0.0078 at $n = 5000$ and 0.0052 at $n = 10000$. Mesner–Shalizi has lower MAE throughout the tested range, decreasing from 0.0290 to 0.0047. The gap between the two estimators narrows from 0.0028 at $n = 5000$ to 0.0004 at $n = 10000$. Although Product-CMI does not produce the smallest MAE in this benchmark, its improvement is most substantial as sample support grows, indicated by the steepest drop in MAE.

Under the evaluated CMIh configuration, pooled MAE remains close to 0.31 across the tested sample sizes. We also evaluate the adaptive-histogram estimator of Marx et al. (2021); its official

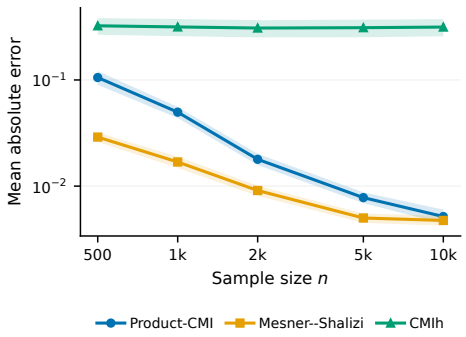


Figure 1: **Estimator accuracy across sample sizes.** Mean absolute error over the full synthetic benchmark (240 runs per point). Shaded regions show 95% bootstrap confidence intervals from 5000 resamples of run-level errors; the vertical axis is logarithmic.

implementation completed only two of the six configurations. To avoid comparing methods over different evaluation sets, the four-method comparison on the common supported subset is reported separately in Appendix C.

5.2 Theory-Guided Finite-Sample Diagnostics

We next examine the three mechanisms emphasized by the analysis: soft-to-hard approximation, continuous dimension, and categorical support.

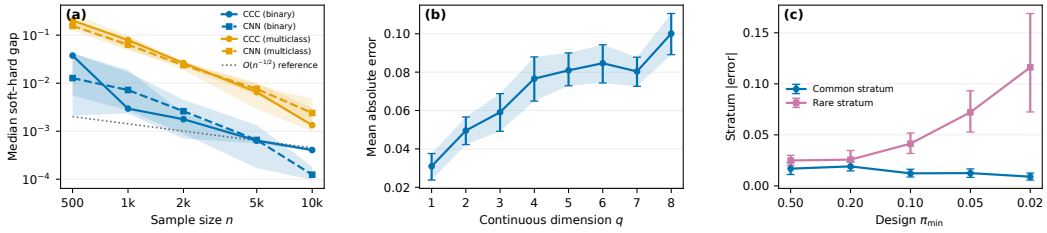


Figure 2: **Theory-guided finite-sample diagnostics.** (a) Median soft–hard gap under $\delta_n = n^{-1/2}$; shaded regions give the interquartile range across twenty seeds, and the dotted line is an $O(n^{-1/2})$ reference. (b) Mean MAE as $q = \dim(W)$ increases at fixed population CMI; intervals are 95% bootstrap confidence intervals across twenty seeds. (c) Mean local absolute error in rare and common categorical strata as rare-stratum probability decreases; intervals are 95% bootstrap confidence intervals across twenty seeds.

Soft-to-hard approximation. Theorem 2 gives an $O_p(a_n)$ soft–hard discrepancy when $\delta_n = O_p(a_n)$. We set $\delta_n = n^{-1/2}$ and evaluate $n = 500$ through 10,000 in two representative configurations: (C, C, C) and (C, N, N), each with binary and multiclass categorical variables under moderate dependence. Hard and soft estimates use the same sample, geometry, and bandwidth within each run. Figure 2a shows a decreasing median gap in all four settings through $n = 10{,}000$. For example, the binary (C, C, C) gap falls from 0.0376 to 4.05×10^{-4}, while the multiclass (C, C, C) gap falls from 0.202 to 1.34×10^{-3}. The $O(n^{-1/2})$ line is included only as a reference slope.

Continuous dimension. To study the continuous-dimension term, we use Gaussian data with conditional correlation $\rho = 0.5$, for which $I(U; V \mid W) = 0.14384$, and vary $q = \dim(W)$ from 1 to 8 at $n = 2000$. The bandwidth is re-selected at each dimension using the same likelihood-CV procedure as in the main benchmark. In Figure 2b, mean MAE rises from 0.0310 at $q = 1$ to 0.0766

at $q = 4$. For $q = 4, \ldots, 8$, it varies non-monotonically between approximately 0.077 and 0.100, with the largest mean MAE occurring at $q = 8$. The initial increase may be interpreted as evidence of increasing finite-sample difficulty, and a monotone dimension effect may not necessarily be exhibited over the larger values tested.

Categorical sparsity. For the categorical-support diagnostic, U and V are numerical and W is binary categorical. We hold the stratum-level population CMI fixed and reduce the rare-stratum probability from 0.50 to 0.02. Figure 2c shows that the mean rare-stratum error increases from 0.0249 to 0.1162, whereas the common-stratum error changes from 0.0169 to 0.0090. This is a local support effect: it does not imply that global CMI error must rise as the stratum becomes rarer, since that stratum also receives less weight in the population average. A fixed-bandwidth version gives the same qualitative rare-stratum pattern, and a like-for-like global Mesner–Shalizi comparison is reported in Appendix C.

5.3 Optimization with a Known Conditional-Independence Solution

We finally test whether soft Product-CMI provides a useful optimization signal. The controlled models satisfy $U \perp V_\theta \mid W \iff \theta = 0$, so the correct solution $\theta^\star = 0$ is known before optimization. The experiment covers all six type configurations, three initial dependence strengths, twenty seeds, and $n = 1000$, for 360 runs. Initial parameter scales are calibrated by configuration; the exact values, parameterizations, and optimizer settings are given in Appendix D. For numerical V, θ enters a smooth structural model, while for categorical V it enters class-specific logits whose probability vectors are passed to soft Product-CMI. Geometry and bandwidth are fixed at their initial-state values.

Only soft Product-CMI is optimized. Hard Product-CMI and Mesner–Shalizi are used only for evaluation. Because the initial scale of θ differs across configurations, we report the normalized residual

$$R_\theta = \frac{|\theta_{\text{final}} - \theta^\star|}{|\theta_{\text{initial}} - \theta^\star|} = \frac{|\theta_{\text{final}}|}{|\theta_{\text{initial}}|},$$

where $R_\theta = 1$ denotes no progress and $R_\theta = 0$ is the known solution. Figure 3a shows $R_\theta < 1$ in all 360 runs. The median is 0.0165, with values ranging from approximately 3×10^{-5} to just below 1. At the configuration level, (C, C, N) has the largest median R_θ (0.137), while the medians for the other five configurations remain below 0.022.

As an independent check, define $\Delta I = I_{\text{initial}} - I_{\text{final}}$. Figure 3b compares ΔI from hard Product-CMI and Mesner–Shalizi for each run. Hard Product-CMI decreases in 358 of 360 runs and is unchanged in two, while Mesner–Shalizi decreases in 356 runs and is unchanged in four; no increases are observed under either estimator. Thus, the reduction in the optimized soft objective is accompanied by non-increasing CMI under two estimators that do not participate in the gradient computation.

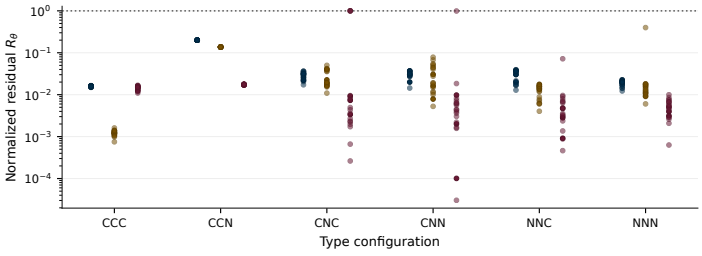


Figure 3: **Optimization toward a known conditional-independence solution.** (a) Normalized residual R_θ after optimizing soft Product-CMI across six type configurations, three initial dependence strengths, and twenty random seeds (360 runs).

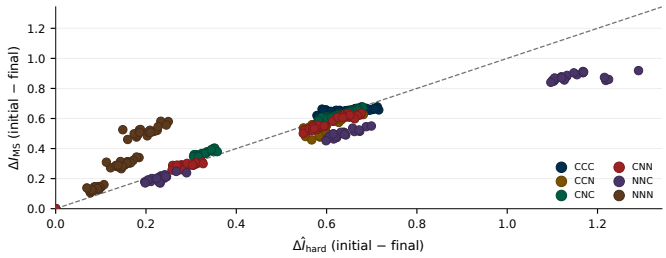


Figure 3: **Optimization toward a known conditional-independence solution (continued).** (b) Changes in hard Product-CMI and Mesner–Shalizi for the same 360 runs; positive values indicate lower estimated CMI after optimization.

A semi-synthetic Adult check also places the Product-CMI repair-grid minimum at the known solution $\beta/\alpha = 1$ in all 30 cases, with no ties. Across the ten seeds, moving from $\beta/\alpha = 0$ to $\beta/\alpha = 1$ reduces mean Product-CMI by 0.4776, 0.6564, and 0.6445 under weak, moderate, and strong dependence, respectively. Details are given in Appendix E.

6 Discussion and Limitations

We introduced Product-CMI, a mixed-type conditional mutual information estimator with a differentiable categorical relaxation. The hard estimator is consistent across all eight categorical–continuous type assignments, and the soft estimator remains within $O_p(a_n)$ of it when the categorical relaxation error is $O_p(a_n)$. Empirically, Product-CMI becomes more accurate as sample support grows, the finite-sample diagnostics follow the mechanisms highlighted by the analysis, and optimizing the soft estimator moves the parameter toward known conditional-independence solutions across all tested configurations. These results support Product-CMI as a differentiable CMI objective for mixed-type optimization without claiming finite-sample estimation dominance.

Good performance of the proposed Product-CMI relies on sufficient local support: rare categorical strata reduce the size of data that can be used, whereas higher continuous dimension makes local smoothing more difficult. Our consistency result is developed for settings with fixed finite categorical support. It is interesting to extend our development to settings in which the number of categories grows with n. The explicit rate in (7) is derived for one representative mixed setting; only the consistency result extends across all eight type assignments.

Product-CMI is not necessarily uniformly superior for finite-sample CMI estimation. Mesner–Shalizi has lower pooled MAE throughout our benchmark, and the gap can be negligible at a large sample size. The main advantage of Product-CMI lies in its mixed-type estimation with a differentiable relaxation whose relation to the hard estimator is controlled statistically.

The soft-transfer theorem (Theorem 2) assumes that the relaxed categorical representations approach their one-hot counterparts and does not guarantee convergence of a particular optimizer or temperature schedule. Our known-optimum experiments also keep geometry and bandwidth fixed at their initial values. The semi-synthetic Adult experiment in Appendix E uses a controlled dependence mechanism with known ground truth and should not be interpreted as evidence for unrestricted data repair or fairness applications.

Finally, finite-sample Product-CMI need not be nonnegative. Before taking logarithms, we floor each local mass at approximately 10^{-12} to avoid numerical instability from very small local masses. The final Product-CMI estimate itself is not clipped or constrained to be nonnegative. Direct evaluation requires $O(n^2)$ time because the local masses are computed from pairwise similarities between observations. Blockwise accumulation reduces memory requirements by processing these pairwise computations in smaller blocks, but it does not reduce the number of pairwise evaluations, so the overall time complexity remains quadratic in n. This can limit the use of the current implementation on larger datasets. Scalable approximations for larger-scale settings are left for future work.

REPRODUCIBILITY STATEMENT

The appendix and supplementary material provide complete proofs for the theoretical results, the synthetic data-generating processes, reference-CMI calculations, preprocessing and bandwidth-selection procedures, baseline configurations, optimization parameterizations, random seeds, and additional experimental results. The implementation of Product-CMI and the scripts used to reproduce the main experiments and figures will be publicly released. The supplementary material further documents the finite-difference check used for the categorical optimization experiment and the semi-synthetic Adult study.

AI USE STATEMENT

Generative AI tools were used to suggest improvements to the clarity and wording of the manuscript. All AI-assisted suggestions were carefully reviewed and revised by the authors to ensure rigor of the material. The authors independently verified all mathematical claims, derivations, citations, implementation choices, and experimental results, and take full responsibility for the final manuscript.

ETHICS STATEMENT

Product-CMI is a general statistical estimation and optimization tool. Its use in individual applications such as fairness, invariant learning, or data repair requires accommodating specific conditional-independence relations to reflect problem-specific features. The semi-synthetic Adult experiment in Appendix E is used only as a methodological evaluation and is not intended as a fairness analysis or as evidence for any particular decision rule.

REFERENCES

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540, 2018.

Bao Duong and Thin Nguyen. Diffeomorphic information neural estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7468–7475, 2023. doi: 10.1609/aaai.v37i6.25908.

Stefan Frenzel and Bernd Pompe. Partial mutual information for coupling analysis of multivariate time series. *Physical Review Letters*, 99(20):204101, 2007. doi: 10.1103/PhysRevLett.99.204101.

Peter Hall, Jeffrey Racine, and Qi Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004. doi: 10.1198/016214504000000548.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004. doi: 10.1103/PhysRevE.69.066138.

Qi Li and Jeffrey S. Racine. Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, 86(2):266–292, 2003. doi: 10.1016/S0047-259X(02)00025-8.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

Alexander Marx, Lincen Yang, and Matthijs van Leeuwen. Estimating conditional mutual information for discrete-continuous mixtures using multidimensional adaptive histograms. In *Proceedings of the 2021 SIAM International Conference on Data Mining*, pp. 387–395, 2021. doi: 10.1137/1.9781611976700.44.

Octavio César Mesner and Cosma Rohilla Shalizi. Conditional mutual information estimation for mixed, discrete and continuous data. *IEEE Transactions on Information Theory*, 67(1):464–484, 2021. doi: 10.1109/TIT.2020.3024886.

Sina Molavipour, Germán Bassi, and Mikael Skoglund. Conditional mutual information neural estimator. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5025–5029, 2020. doi: 10.1109/ICASSP40776.2020.9053422.

Arnab Mondal, Arnab Bhattacharjee, Sudipto Mukherjee, Himanshu Asnani, Sreeram Kannan, and Prathosh A P. C-mi-gan: Estimation of conditional mutual information using minmax formulation. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 849–858, 2020.

Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. Ccmi: Classifier-based conditional mutual information estimation. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, volume 115 of *Proceedings of Machine Learning Research*, pp. 1083–1093, 2020.

A. Tuan Nguyen, Philip H. S. Torr, and Ser-Nam Lim. FedSR: A simple and effective domain generalization method for federated learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38831–38843, 2022.

Alireza Pirhadi, Mohammad Hossein Moslemi, Alexander Cloninger, Mostafa Milani, and Babak Salimi. OTClean: Data cleaning for conditional independence violations using optimal transport. *Proceedings of the ACM on Management of Data*, 2(3):1–26, 2024. doi: 10.1145/3654963.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180, 2019.

Oana-Iuliana Popescu, Andreas Gerhardus, Martin Rabel, and Jakob Runge. Non-parametric conditional independence testing for mixed continuous-categorical variables: A novel method and numerical evaluation. In *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pp. 406–450, 2025.

Jeffrey Racine and Qi Li. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1):99–130, 2004. doi: 10.1016/S0304-4076(03)00157-X.

Lei Zan, Anouar Meynaoui, Charles K. Assaad, Emilie Devijver, and Eric Gaussier. A conditional mutual information estimator for mixed data and an associated conditional independence test. *Entropy*, 24(9):1234, 2022. doi: 10.3390/e24091234.

SUPPLEMENTARY MATERIAL

A COMPLETE PROOFS

A.1 PROOF OF LEMMA 1

Fix $S \in \mathcal{S}$ and write $X_S = (X_S^{\mathrm{cat}}, X_S^{\mathrm{con}})$, with $X_S^{\mathrm{con}} \in \mathbb{R}^{d_S}$.

The hard Product local mass is

$$M_{S,i}^{\mathrm{H}}(h_n) = \frac{1}{n-1} \sum_{j \neq i} K_{S,ij}^{\mathrm{H}}(h_n).$$

Suppose first that $d_S > 0$. Conditional on $X_{S,i} = x_S$,

$$\mathbb{E}\big[M_{S,i}^{\mathrm{H}}(h_n) \mid X_{S,i} = x_S\big] = \int_{\mathbb{R}^{d_S}} \exp\left(-\frac{\|x_S^{\mathrm{con}} - t\|_2^2}{2h_n^2}\right) g_S(x_S^{\mathrm{cat}}, t)\, dt.$$

Using $t = x_S^{\mathrm{con}} + h_n z$ and $dt = h_n^{d_S}\, dz$ gives

$$\mathbb{E}\big[M_{S,i}^{\mathrm{H}}(h_n) \mid X_{S,i} = x_S\big] = h_n^{d_S} \int_{\mathbb{R}^{d_S}} e^{-\|z\|_2^2/2} g_S(x_S^{\mathrm{cat}}, x_S^{\mathrm{con}} + h_n z)\, dz.$$

A second-order Taylor expansion along the continuous coordinates yields

$$g_S(x_S^{\mathrm{cat}}, x_S^{\mathrm{con}} + h_n z) = g_S(x_S) + h_n \nabla g_S(x_S)^\top z + O(h_n^2 \|z\|_2^2).$$

The linear term integrates to zero by Gaussian symmetry. Since

$$\int_{\mathbb{R}^{d_S}} e^{-\|z\|_2^2/2}\, dz = (2\pi)^{d_S/2},$$

we obtain

$$\mathbb{E}\big[M_{S,i}^{\mathrm{H}}(h_n) \mid X_{S,i}\big] = b_S(h_n)\left[g_S(X_{S,i}) + O(h_n^2)\right]. \tag{10}$$

If $d_S = 0$, every coordinate in S is categorical and

$$M_{S,i}^{\mathrm{H}} = \frac{1}{n-1} \sum_{j \neq i} \mathbf{1}\{X_{S,j} = X_{S,i}\}.$$

Therefore

$$\mathbb{E}[M_{S,i}^{\mathrm{H}} \mid X_{S,i} = x_S] = g_S(x_S),$$

which agrees with (10) because $b_S = 1$ and there is no smoothing bias.

We next control the stochastic fluctuation. Conditional on $X_{S,i}$, the summands are independent, bounded by one, and have conditional expectation and variance of order $b_S(h_n)$. Bernstein's inequality gives, for fixed i,

$$M_{S,i}^{\mathrm{H}} - \mathbb{E}[M_{S,i}^{\mathrm{H}} \mid X_{S,i}] = O_p\left(\sqrt{\frac{b_S(h_n)t}{n}} + \frac{t}{n}\right).$$

Taking $t = c \log n$ with $c > 1$ and applying a union bound over $i = 1, \ldots, n$ gives

$$\max_i \left|M_{S,i}^{\mathrm{H}} - \mathbb{E}[M_{S,i}^{\mathrm{H}} \mid X_{S,i}]\right| = O_p\left(\sqrt{\frac{b_S(h_n)\log n}{n}} + \frac{\log n}{n}\right). \tag{11}$$

When $d_S = 0$, the right-hand side is $o_p(1)$ directly. When $d_S > 0$, $b_S(h_n) \asymp h_n^{d_S}$. Because $d_S \leq D$ and $h_n < 1$ eventually, $h_n^{d_S} \geq h_n^D$. Thus A5 implies $n h_n^{d_S} / \log n \to \infty$, and the stochastic error in (11) is $o_p(b_S(h_n))$.

Combining the bias and stochastic terms,

$$M_{S,i}^{\mathrm{H}}(h_n) = b_S(h_n) g_S(X_{S,i})\{1 + o_p(1)\}, \tag{12}$$

uniformly over the sample evaluation points and simultaneously over $S \in \mathcal{S}$. This proves Lemma 1.

A.2 Proof of Theorem 1

Applying (12) to the four local masses gives

$$\begin{aligned} M^{\mathrm{H}}_{UVW,i} &= b_{UVW} g_{UVW}(U_i, V_i, W_i)\{1 + o_p(1)\}, \\ M^{\mathrm{H}}_{UW,i} &= b_{UW} g_{UW}(U_i, W_i)\{1 + o_p(1)\}, \\ M^{\mathrm{H}}_{VW,i} &= b_{VW} g_{VW}(V_i, W_i)\{1 + o_p(1)\}, \\ M^{\mathrm{H}}_{W,i} &= b_{W} g_{W}(W_i)\{1 + o_p(1)\}. \end{aligned}$$

Substituting these expressions into the local CMI ratio,

$$\widehat{r}^{\mathrm{H}}_i = \frac{b_{UVW}(h_n) b_W(h_n)}{b_{UW}(h_n) b_{VW}(h_n)} r(U_i, V_i, W_i)\{1 + o_p(1)\}.$$

The continuous dimensions satisfy

$$d_{UVW} + d_W = (d_U + d_V + d_W) + d_W = (d_U + d_W) + (d_V + d_W) = d_{UW} + d_{VW}.$$

Therefore,

$$b_{UVW}(h_n) b_W(h_n) = b_{UW}(h_n) b_{VW}(h_n).$$

The deterministic kernel-volume terms cancel exactly, leaving

$$\widehat{r}^{\mathrm{H}}_i = r(U_i, V_i, W_i)\{1 + o_p(1)\}$$

uniformly over i.

By A3 and (12), the four local masses are positive with probability tending to one. The logarithm is therefore well defined asymptotically, and

$$\max_i \left| \log \widehat{r}^{\mathrm{H}}_i - \log r(U_i, V_i, W_i) \right| \xrightarrow{p} 0. \tag{13}$$

Now decompose

$$\begin{aligned} \widehat{I}^{\mathrm{H}}_n - I(U; V \mid W) = \frac{1}{n} \sum_{i=1}^{n} \left[\log \widehat{r}^{\mathrm{H}}_i - \log r(U_i, V_i, W_i) \right] \\ + \left[\frac{1}{n} \sum_{i=1}^{n} \log r(U_i, V_i, W_i) - \mathbb{E}[\log r(U, V, W)] \right]. \end{aligned}$$

The first term converges to zero by (13). The second converges to zero by the law of large numbers and A4. Hence

$$\widehat{I}^{\mathrm{H}}_n \xrightarrow{p} I(U; V \mid W).$$

The proof depends on the variable types only through their continuous dimensions d_X. It therefore covers all eight categorical–continuous assignments of (U, V, W), proving Theorem 1.

A.3 Derivation of Expression (7)

Consider the setting of Section 4.2, where U and W are categorical and $V \in \mathbb{R}^d$. Let

$$\pi_{\min} = \min_{u,w:\mathbb{P}(U=u,W=w)>0} \mathbb{P}(U = u, W = w).$$

The second-order expansion above contributes smoothing bias $O(h_n^2)$.

Within a categorical stratum (u, w), the expected local kernel mass has scale $\pi_{uw} h_n^d$. The same Bernstein argument therefore gives relative local stochastic error

$$O_p\left(\sqrt{\frac{\log n}{n \pi_{\min} h_n^d}} + \frac{\log n}{n \pi_{\min} h_n^d} \right). \tag{14}$$

Under $n\pi_{\min}h_n^d/\log n \to \infty$, the second term in (14) is asymptotically smaller than the first.

If, in addition, $\mathbb{E}[(\log r(U,V,W))^2] < \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n} \log r(U_i, V_i, W_i) - I(U;V \mid W) = O_p(n^{-1/2}).$$

Combining the smoothing, local stochastic, and empirical-averaging terms,

$$\widehat{I}_n^{\mathrm{H}} - I(U;V \mid W) = O_p\left(h_n^2 + \sqrt{\frac{\log n}{n\pi_{\min}h_n^d}} + n^{-1/2}\right),$$

which is (7). This is a diagnostic first-order error scale rather than a sharp or minimax-optimal rate.

A.4 Proof of Theorem 2

For a categorical variable X, let e_{X_i} denote its one-hot representation and let p_i^X denote the corresponding soft probability vector. Recall

$$\delta_n = \max_{\substack{X\in\{U,V,W\}\\ X \text{ categorical}}} \max_i \|p_i^X - e_{X_i}\|_1.$$

For one categorical coordinate,

$$\begin{aligned}\left|(p_i^X)^\top p_j^X - e_{X_i}^\top e_{X_j}\right| &\le \left|(p_i^X - e_{X_i})^\top p_j^X\right| + \left|e_{X_i}^\top (p_j^X - e_{X_j})\right| \\ &\le 2\delta_n.\end{aligned}$$

If a Product factor contains $m \le 3$ categorical coordinates, then, because all hard and soft similarities lie in $[0,1]$,

$$\left|\prod_{\ell=1}^{m} S_{\ell,ij} - \prod_{\ell=1}^{m} H_{\ell,ij}\right| \le \sum_{\ell=1}^{m} |S_{\ell,ij} - H_{\ell,ij}| \le 6\delta_n.$$

For $S \in \mathcal{S}$, define the product of its continuous Gaussian factors by

$$G_{S,ij} = \prod_{\substack{X\in S\\ X \text{ continuous}}} \exp\left(-\frac{\|X_i - X_j\|_2^2}{2h_n^2}\right),$$

with $G_{S,ij} = 1$ when S contains no continuous coordinate.

Then

$$\left|M_{S,i}^{\mathrm{S}} - M_{S,i}^{\mathrm{H}}\right| \le 6\delta_n \left[\frac{1}{n-1}\sum_{j\neq i} G_{S,ij}\right].$$

The bracketed term is an ordinary d_S-dimensional Gaussian local mass without categorical restriction. The same kernel expectation and concentration calculation gives

$$\max_i \frac{1}{b_S(h_n)}\left[\frac{1}{n-1}\sum_{j\neq i} G_{S,ij}\right] = O_p(1).$$

When $d_S = 0$, the bracketed term is identically one.

Therefore,

$$\max_i \left|M_{S,i}^{\mathrm{S}} - M_{S,i}^{\mathrm{H}}\right| = O_p\left(\delta_n b_S(h_n)\right). \tag{15}$$

By (12) and A3, $M_{S,i}^{\mathrm{H}} \asymp b_S(h_n)$ uniformly with probability tending to one. Dividing (15) by the hard local mass gives

$$\max_i \left|\frac{M_{S,i}^{\mathrm{S}}}{M_{S,i}^{\mathrm{H}}} - 1\right| = O_p(\delta_n). \tag{16}$$

Suppose now that $\delta_n = O_p(a_n)$ and $a_n \to 0$. Applying (16) to each of the four masses gives

$$M^{\mathrm{S}}_{S,i} = M^{\mathrm{H}}_{S,i}\{1 + O_p(a_n)\}$$

uniformly over i.

Hence

$$\log \widehat{r}^{\mathrm{S}}_i - \log \widehat{r}^{\mathrm{H}}_i = \log \frac{M^{\mathrm{S}}_{UVW,i}}{M^{\mathrm{H}}_{UVW,i}} + \log \frac{M^{\mathrm{S}}_{W,i}}{M^{\mathrm{H}}_{W,i}} - \log \frac{M^{\mathrm{S}}_{UW,i}}{M^{\mathrm{H}}_{UW,i}} - \log \frac{M^{\mathrm{S}}_{VW,i}}{M^{\mathrm{H}}_{VW,i}}.$$

Each ratio on the right is $1 + O_p(a_n)$ uniformly. Since $|\log(1 + x)| \le 2|x|$ for sufficiently small $|x|$,

$$\max_i \left|\log \widehat{r}^{\mathrm{S}}_i - \log \widehat{r}^{\mathrm{H}}_i\right| = O_p(a_n).$$

Averaging over i yields

$$\widehat{I}^{\mathrm{S}}_n - \widehat{I}^{\mathrm{H}}_n = O_p(a_n).$$

Together with Theorem 1,

$$\widehat{I}^{\mathrm{S}}_n \xrightarrow{p} I(U; V \mid W),$$

which proves Theorem 2.

In the representative mixed setting,

$$\widehat{I}^{\mathrm{S}}_n - I(U; V \mid W) = O_p\left(h_n^2 + \sqrt{\frac{\log n}{n\pi_{\min} h_n^d}} + n^{-1/2} + a_n\right).$$

B Experimental Setup and Reproducibility Details

B.1 Synthetic benchmark

All synthetic arrays use the column order $(U, V, W) = (0, 1, 2)$, and the target is always $I(U; V \mid W)$.

The benchmark uses six estimator type configurations,

$$(C, C, C),\ (C, C, N),\ (C, N, C),\ (C, N, N),\ (N, N, C),\ (N, N, N),$$

which cover the eight assignments up to the U–V symmetry of CMI. Here C denotes categorical treatment and N denotes numerical treatment by the estimator.

For each sample size, the benchmark crosses six configurations, two generative families, four dependence levels, and five seeds: $6 \times 2 \times 4 \times 5 = 240$ runs. The final sample-size grid is $n \in \{500, 1000, 2000, 5000, 10000\}$.

Binary family. We generate $U_i \sim \mathrm{Bernoulli}(0.5)$ and $W_i \sim \mathrm{Bernoulli}(0.5)$ independently, and

$$V_i = W_i + \beta_U U_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, 0.8^2).$$

The dependence coefficient is $\beta_U = 1.5s$, with

level	null	weak	moderate	strong
s	0	0.25	0.55	0.90
β_U	0	0.375	0.825	1.35

Multiclass family. We use

$$\mathbb{P}(U = k) = (0.45, 0.25, 0.20, 0.10)_k, \qquad \mathbb{P}(W = j) = (0.40, 0.35, 0.25)_j,$$

with category effects

$$u_{\text{eff}} = (-1.5, -0.5, 0.5, 1.5), \qquad w_{\text{eff}} = (-1, 0, 1).$$

The numerical variable is generated as

$$\mu_i = w_{\text{eff}}[W_i] + \beta\, u_{\text{eff}}[U_i],$$
$$V_i \mid U_i, W_i \sim \mathcal{N}(\mu_i, 0.8^2).$$

The fixed dependence coefficients are

level	null	weak	moderate	strong
β	0	0.1870574951	0.4085731506	0.6663780212

Categorical V. When V is treated categorically, the generated numerical V is discretized into five bins. For each family, the bin edges are fixed before benchmark evaluation using $500{,}000$ samples generated under the strong-dependence setting with seed 20260917.

If $e_1, \ldots, e_4$ denote the empirical quantiles at probabilities $0.2,\ 0.4,\ 0.6,\ 0.8$, then

$$V_{b,i} = \text{clip}\left(\sum_{k=1}^{4} \mathbf{1}\{V_i \geq e_k\}, 0, 4\right).$$

The same fixed binning rule is used when computing the corresponding reference-CMI values.

Table 1: Estimator treatment used for the six benchmark configurations.

ID	(U, V, W)	Categorical coordinates	Numerical coordinates	Treatment of V
A_CCC	(C, C, C)	U, V, W	–	five-bin discretization
B_CCN	(C, C, N)	U, V	W	five-bin discretization
C_CNC	(C, N, C)	U, W	V	MinMax scaling
D_CNN	(C, N, N)	U	V, W	MinMax scaling
E_NNC	(N, N, C)	W	U, V	MinMax scaling
F_NNN	(N, N, N)	–	U, V, W	MinMax scaling

Estimator representation. In the validated Product-CMI implementation, MinMax scaling is applied to the generated V column when V is treated numerically. Numerically treated U or W coordinates retain their generator scale. Mesner–Shalizi is evaluated on the same transformed tensor as Product-CMI. CMIh and Marx use the corresponding raw benchmark arrays, with categorical V binned using the same fixed edges.

The numerical treatment used in these empirical configurations is an implementation convention for the benchmark and is distinct from the continuous-distribution assumptions used in the theoretical analysis.

Reference CMI. Reference values are computed independently of the samples supplied to the estimators using Monte Carlo samples of size $500{,}000$. For numerical V, the reference calculation uses the model likelihood ratio. For binned V, it uses the corresponding empirical discrete CMI with the fixed binning rule above. The reference values are fixed before estimator evaluation and are never used for bandwidth selection or hyperparameter tuning.

B.2 Bandwidth selection

Product-CMI uses one scalar continuous bandwidth shared by the Gaussian factors entering its local masses.

Let ϕ_{con} denote the continuous representation under the fitted geometry. The reference scale is

$$\sigma_{\mathrm{con}}^{(\mathrm{ref})} = \max\left\{\mathrm{median}_{i<j}\,\|\phi_{\mathrm{con},i} - \phi_{\mathrm{con},j}\|_2, 10^{-6}\right\}.$$

The candidate bandwidths are

$$\sigma_{\mathrm{cand}} = \mathrm{clip}\left(\sigma_{\mathrm{con}}^{(\mathrm{ref})} f, 10^{-3}, 10^{3}\right), \qquad f \in \left\{\frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4\right\}.$$

The selected value maximizes five-fold held-out likelihood. The folds are shuffled with seed $42 +$ dataset seed. Bandwidth selection and CMI evaluation use the same fitted geometry, and reference CMI is not used in the selection procedure.

For the optimization experiments, geometry and bandwidth are fitted once at the initial state and then held fixed.

B.3 Numerical safeguard and baseline implementations

Before taking logarithms, local masses are floored at $\max\{\epsilon_{\mathrm{machine}}, 10^{-12}\}$. The final Product-CMI estimate itself is not clipped.

The baseline settings are summarized in Table 2.

Table 2: Baseline implementations used in the estimator benchmark.

Method	Implementation	Setting
Mesner–Shalizi	Tigramite `CMIknnMixed`	$k = 7$, global, MS estimator, no transform
CMIh	official `mixedEstimator()`	$k \approx 0.1\times$ stratum size; output in nats
Marx et al.	official adaptive-histogram implementation	`eps` $= 0$, `Kmax` $= 0$

The Mesner–Shalizi runs use Tigramite version 5.2.10.1. The validated CMIh code corresponds to commit `efe407c`, with only a compatibility wrapper change from `rowMax` to `rowMaxs`. The Marx implementation corresponds to commit `264e65e`.

The official Marx implementation completed the A_CCC and C_CNC configurations but did not complete the remaining four configurations. Consequently, Marx is not pooled with the full six-configuration benchmark.

B.4 Evaluation metric and uncertainty

Estimator accuracy is measured by run-level absolute error,

$$\left|\widehat{I} - I_{\mathrm{ref}}\right|,$$

in nats.

For the sample-size benchmark, confidence intervals are computed separately for each method and sample size by resampling the run-level absolute errors with replacement 5000 times. The reported 95% interval is the 2.5th–97.5th percentile interval of the bootstrap means. The bootstrap seed is 42.

For the continuous-dimension diagnostic, the bootstrap unit is one of the twenty seed-level errors. For the sparsity diagnostic, the bootstrap unit is one of the twenty seed-level stratum errors. Both use the same 5000-resample, seed-42 percentile procedure. The soft-to-hard panel reports medians and interquartile ranges rather than bootstrap intervals.

C Additional Experimental Results

C.1 Sample-size benchmark

For completeness, Table 3 reports the pooled MAE values underlying the main sample-size comparison.

Table 3: Pooled mean absolute error relative to independently computed reference CMI.

n	Product-CMI	Mesner–Shalizi	CMIh
500	0.1054	0.0290	0.3238
1000	0.0497	0.0169	0.3159
2000	0.0179	0.0091	0.3084
5000	0.0078	0.0050	0.3108
10000	0.00517	0.00475	0.3150

At $n = 10000$, the corresponding 95% bootstrap confidence intervals are $[0.00442, 0.00601]$ for Product-CMI, $[0.00424, 0.00527]$ for Mesner–Shalizi, and $[0.258, 0.374]$ for CMIh.

Marx common-supported subset. The four-method comparison is evaluated separately on the common subset supported by the official Marx implementation: A_CCC and C_CNC. This subset contains $2 \times 4 \times 2 \times 5 = 80$ runs for each reported sample size, as shown in Table 4.

Table 4: Mean absolute error on the common subset supported by all four estimators (A_CCC and C_CNC). Lower is better.

n	Product-CMI	Mesner–Shalizi	CMIh	Marx
500	0.0847	0.0267	0.0232	0.0279
1000	0.0375	0.0167	0.0135	0.0163
2000	0.0127	0.0084	0.0045	0.0079
5000	0.0040	0.0046	0.0038	0.0050

C.2 Soft-to-hard approximation

The soft-to-hard diagnostic uses A_CCC and D_CNN under moderate dependence, with binary and multiclass families, seeds 0–19, and $n \in \{500, 1000, 2000, 5000, 10000\}$. Within each run, the hard and soft estimators use the same generated sample, geometry, and bandwidth. The main relaxation is $\delta_n = n^{-1/2}$.

At $n = 10000$, the median absolute soft–hard gaps are 4.05×10^{-4} and 1.25×10^{-4} for the binary A_CCC and D_CNN settings, respectively. The corresponding multiclass values are 1.34×10^{-3} and 2.39×10^{-3}.

The uncertainty bands in Figure 2(a) are interquartile ranges across the twenty seeds. The displayed $O(n^{-1/2})$ line is a reference slope only.

C.3 Continuous dimension

The continuous-dimension diagnostic uses Gaussian data with fixed conditional correlation $\rho = 0.5$, so that

$$I_{\text{true}} = -\frac{1}{2}\log(1 - \rho^2) = 0.1438410362.$$

We use $n = 2000$, $q = \dim(W) \in \{1, \ldots, 8\}$, with twenty seeds. The bandwidth is selected independently for each generated dataset using the same likelihood-CV procedure as in the main benchmark. The resulting MAE values are reported in Table 5.

Table 5: Product-CMI MAE as the continuous dimension increases. Confidence intervals bootstrap the twenty seed-level errors using 5000 resamples and seed 42.

q	Mean MAE	95% CI
1	0.031	[0.024, 0.038]
2	0.050	[0.042, 0.057]
3	0.059	[0.049, 0.069]
4	0.077	[0.065, 0.088]
5	0.081	[0.073, 0.090]
6	0.085	[0.074, 0.094]
7	0.080	[0.073, 0.088]
8	0.100	[0.089, 0.111]

C.4 Categorical sparsity

The sparsity diagnostic is separate from the six-configuration benchmark. It uses $n = 2000$, $\rho = 0.5$, with numerical U and V and binary categorical W. The conditional correlation, and hence the stratum-level population CMI, is held fixed while the designed rare-stratum probability varies over $\pi \in \{0.50, 0.20, 0.10, 0.05, 0.02\}$.

For each π, twenty random seeds are used. If η_i denotes the pointwise Product-CMI log-ratio contribution, the stratum-level diagnostic is

$$\left| \frac{1}{|S|} \sum_{i \in S} \eta_i - I_{\text{true}} \right|.$$

The corresponding rare- and common-stratum errors are reported in Table 6.

Table 6: Local Product-CMI error under decreasing categorical support. Values are means over twenty seeds.

π	Rare-stratum error	Common-stratum error
0.50	0.0249	0.0169
0.20	0.0257	0.0191
0.10	0.0414	0.0123
0.05	0.0721	0.0125
0.02	0.1162	0.0090

The main sparsity panel uses a cross-validated bandwidth for each dataset. We also repeated the diagnostic with a fixed bandwidth calibrated from the balanced $\pi = 0.50$ setting using seeds 0–9 and then held fixed as π decreased. The rare-stratum deterioration remains under this fixed-bandwidth check, so the effect is not explained solely by bandwidth adaptation.

Mesner–Shalizi with fixed $k = 7$ was evaluated on the same sparsity grid using global CMI error. Its global error is non-monotone in π and does not reproduce the monotone rare-stratum pattern above. We do not report a rare/common decomposition for Mesner–Shalizi because that diagnostic is defined from the pointwise Product-CMI contributions.

D Known-Optimum Optimization Details

The known-optimum experiment is constructed so that

$$U \perp V_\theta \mid W \quad \Longleftrightarrow \quad \theta = 0,$$

so that $\theta^\star = 0$ is known before optimization.

The study uses all six type configurations, three initial dependence strengths, twenty seeds, 0–19, and $n = 1000$, for a total of 360 optimization runs.

The initial values are

Table 7: Initial dependence parameter used in the known-optimum experiment.

Configuration	Weak	Moderate	Strong
(C, C, C)	0.25	0.50	1.00
(C, C, N)	0.0625	0.125	0.25
(C, N, C)	0.50	1.00	2.00
(C, N, N)	0.50	1.00	2.00
(N, N, C)	0.50	1.00	2.00
(N, N, N)	0.50	1.00	2.00

For numerical V, θ enters a smooth structural dependence model. For categorical V, θ enters class-specific logits; the resulting probability vectors are passed directly to soft Product-CMI. The final categorical parameterization uses scale $s = 4$ and temperature $\tau = 2$.

Only soft Product-CMI contributes to the optimization gradient. Hard Product-CMI and Mesner–Shalizi are evaluated only at the initial and final states.

We use Adam with learning rate 0.05 for at most 100 iterations. Optimization stops when either $|\theta| < 0.02$ or the soft Product-CMI objective fails to improve by more than 10^{-7} for 15 consecutive iterations. Geometry and bandwidth are fitted once at the initial state and then held fixed; bandwidth selection is not part of the computational graph.

For representative categorical cases, automatic differentiation is compared with a central finite-difference approximation using $\varepsilon = 10^{-3}$ at $\theta = \pm 0.5$. The relative discrepancy is approximately 5×10^{-6}.

E Semi-Synthetic Adult Study

The Adult experiment is a controlled semi-synthetic check using real covariates while retaining a known conditional-independence solution.

We use $n = 1500$, $U =$ sex, $W =$ education-num from Adult (Becker & Kohavi, 1996). The education variable is standardized. For each seed, we construct

$$V_{\text{clean}} = W_{\text{std}} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 0.35^2).$$

With binary encoding $U \in \{0, 1\}$, define $g(U) = 2U - 1$.

Controlled dependence is introduced through

$$V_{\text{corrupt}} = V_{\text{clean}} + \alpha g(U),$$

where $\alpha \in \{0.5, 1, 2\}$ corresponds to the weak, moderate, and strong settings.

We evaluate the restricted repair family

$$V_{\text{repair}}(\beta) = V_{\text{corrupt}} - \beta g(U).$$

Within this family, the correct repair is $\beta^\star = \alpha$.

For each corruption strength and each of ten seeds, Product-CMI is evaluated over $\frac{\beta}{\alpha} \in \{0, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50\}$. The geometry and bandwidth are fitted from the corrupted initial state and then held fixed while evaluating the repair grid.

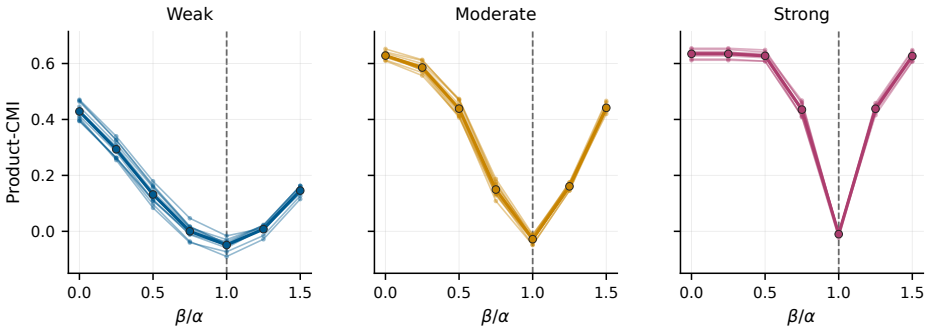


Figure 4: **Semi-synthetic Adult objective curves.** Product-CMI as a function of the normalized repair parameter β/α under weak, moderate, and strong injected dependence. The dashed vertical line marks the known solution $\beta/\alpha = 1$. Across the ten random seeds for each setting, the empirical minimum occurs at the correct repair value in all 30 seed–strength combinations.