A Proof of optimizing a CMI query

Proof of Lemma 1. We use the product-sum property of the logarithm (line 3) and linearity of expectation (line 4) to show that CrossCat's variable partition $\boldsymbol{\gamma}$ induces a factorization of a CMI query.

$$\begin{aligned}
\mathcal{I}_{\mathcal{G}}\left(\boldsymbol{x}_{\mathcal{A}}{:}\boldsymbol{x}_{\mathcal{B}}|\hat{\boldsymbol{x}}_{\mathcal{C}}\right) &= \mathbb{E}\left[\log\left(\frac{p_{\mathcal{G}}(\boldsymbol{x}_{\mathcal{A}}{:}\boldsymbol{x}_{\mathcal{B}}|\hat{\boldsymbol{x}}_{\mathcal{C}})}{p_{\mathcal{G}}(\boldsymbol{x}_{\mathcal{A}}|\hat{\boldsymbol{x}}_{\mathcal{C}})p_{\mathcal{G}}(\boldsymbol{x}_{\mathcal{B}}|\hat{\boldsymbol{x}}_{\mathcal{C}})}\right)\right] \\
&= \mathbb{E}\left[\log\left(\prod_{\mathcal{V}\in\gamma}\frac{p_{\mathcal{G}_{\mathcal{V}}}(\boldsymbol{x}_{\mathcal{A}\cap\mathcal{V}},\boldsymbol{x}_{\mathcal{B}\cap\mathcal{V}}|\hat{\boldsymbol{x}}_{\mathcal{C}\cap\mathcal{V}})}{p_{\mathcal{G}_{\mathcal{V}}}(\boldsymbol{x}_{\mathcal{A}\cap\mathcal{V}}|\hat{\boldsymbol{x}}_{\mathcal{C}\cap\mathcal{V}})p_{\mathcal{G}_{\mathcal{V}}}(\boldsymbol{x}_{\mathcal{B}\cap\mathcal{V}}|\hat{\boldsymbol{x}}_{\mathcal{C}\cap\mathcal{V}})}\right)\right] \\
&= \mathbb{E}\left[\sum_{\mathcal{V}\in\gamma}\log\left(\frac{p_{\mathcal{G}_{\mathcal{V}}}(\boldsymbol{x}_{\mathcal{A}\cap\mathcal{V}},\boldsymbol{x}_{\mathcal{B}\cap\mathcal{V}}|\hat{\boldsymbol{x}}_{\mathcal{C}\cap\mathcal{V}})}{p_{\mathcal{G}_{\mathcal{V}}}(\boldsymbol{x}_{\mathcal{A}\cap\mathcal{V}}|\hat{\boldsymbol{x}}_{\mathcal{C}\cap\mathcal{V}})p_{\mathcal{G}_{\mathcal{V}}}(\boldsymbol{x}_{\mathcal{B}\cap\mathcal{V}}|\hat{\boldsymbol{x}}_{\mathcal{C}\cap\mathcal{V}})}\right)\right] \\
&= \sum_{\mathcal{V}\in\gamma}\mathbb{E}\left[\log\left(\frac{p_{\mathcal{G}_{\mathcal{V}}}(\boldsymbol{x}_{\mathcal{A}\cap\mathcal{V}},\boldsymbol{x}_{\mathcal{B}\cap\mathcal{V}}|\hat{\boldsymbol{x}}_{\mathcal{C}\cap\mathcal{V}})}{p_{\mathcal{G}_{\mathcal{V}}}(\boldsymbol{x}_{\mathcal{A}\cap\mathcal{V}}|\hat{\boldsymbol{x}}_{\mathcal{C}\cap\mathcal{V}})p_{\mathcal{G}_{\mathcal{V}}}(\boldsymbol{x}_{\mathcal{B}\cap\mathcal{V}}|\hat{\boldsymbol{x}}_{\mathcal{C}\cap\mathcal{V}})}\right)\right] \\
&= \sum_{\mathcal{V}\in\gamma}\mathcal{I}_{\mathcal{G}_{\mathcal{V}}}\left(\boldsymbol{x}_{\mathcal{A}\cap\mathcal{V}}{:}\boldsymbol{x}_{\mathcal{B}\cap\mathcal{V}}|\hat{\boldsymbol{x}}_{\mathcal{C}\cap\mathcal{V}}\right).
\end{aligned}$$

B Experimental methods for dependence detection baselines

In this section we outline the methodology used to produce the pairwise R^2 and HSIC heatmaps shown in Figures 6a and 6b. To detect the strength of linear correlation (for R^2) and perform a marginal independence test (for HSIC) given variables x_i and x_j in the Gapminder dataset, all records in which at least one of these two variables is missing were dropped. If the total number of remaining observations was less than three, the null hypothesis of independence was not rejected due to degeneracy of these methods at very small sample sizes. Hypothesis tests were performed at the $\alpha = 0.05$ significance level. To account for multiple testing (a total of $\binom{320}{2} = 51040$), a standard Bonferroni correction was applied to ensure a family-wise error rate of at most α.

We used an open source MATLAB implementation for HSIC (function `hsicTestBoot` from `http://gatsby.ucl.ac.uk/~gretton/indepTestFiles/indep.htm`). 1000 permutations were used to approximate the null distribution, and kernel sizes were determined using median distances from the dataset. From Figure 6b, HSIC detects a large number of statistically significant dependencies. Figures 8 and 9 report spurious relationships reported as dependent by HSIC but have a low dependence probability of less than 0.15 according to posterior CMI (Eq 7), and common-sense relationships reported as independent HSIC but have a high dependence probability.

Figure 8: Spurious relationships detected as dependent by HSIC ($p \ll 10^{-6}$) but probably independent ($\mathbb{P}[\mathcal{I}_{\mathcal{G}}(x_i{:}x_j) > 0] < 0.15$) by the MI upper bound.

variable A	variable B
body mass index (men)	privately owned forest (%)
energy use (per capita)	50+ yrs sex ratio
homicide (15-29)	inflation (annual %)
children out of primary school	male above 60 (%)
income share (fourth 20%)	suicide rate age 45-59
personal computers	arms exports (US$)
billionaires (per 1M)	mobile subscription (per 100)
residential elec. consumption	smear-positive detection (%)
coal consumption (per capita)	age at 1st marriage (women)
coal consumption (per capita)	dead kids per woman
underweight children	murder (per 100K)
female 0-4 years (%)	15+ literacy rate (%)
broadband subscribers (%)	dots new case detection (%)
crude oil prod. (per capita)	TB incidence (per 100K)
dependency ratio	people living with hiv

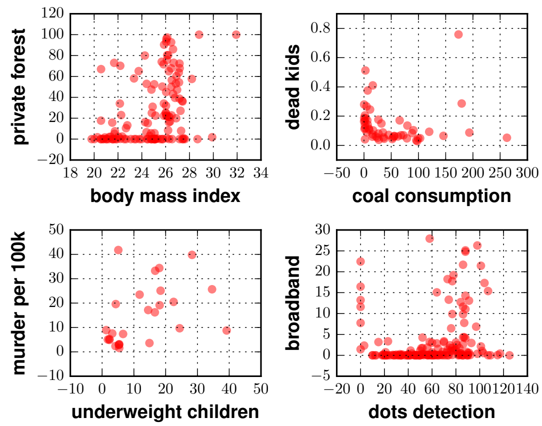
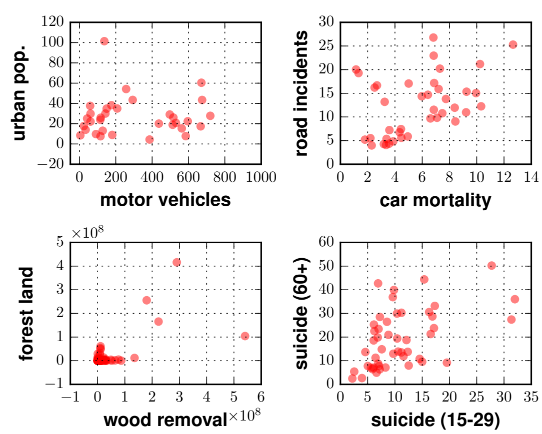


Figure 9: Common-sense relationships detected as independent by HSIC ($p \ll 10^{-6}$), but probably dependent ($\mathbb{P}[\mathcal{I}_{\mathcal{G}}(x_i{:}x_j) > 0] > 0.85$) by the MI upper bound.

variable A	variable B
motor vehicles per 1k	pop urban agglomerations (%)
car mortality (per 100K)	road incidents 45-59
wood removal (cubic m.)	primary forest land (ha)
forest products total (US$)	primary forest land (ha)
15-24 yrs sex ratio	50+ yrs sex ratio
gdp per working hr (US$)	urban population (%)
dots pop. coverage (%)	dots new case detection (%)
total 15-24 unemp. (%)	long term unemp. (%)
female self-employed (%)	female service workers (%)
female agricult. workers (%)	total industry workers (%)
female industry workers (%)	male industry workers (%)
road incidents age 60+	road incidents age 15-29
suicide rate age 15-29	suicide rate age 60+



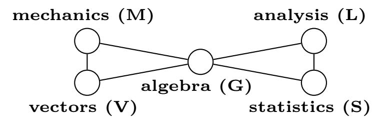
C Application to a database of mathematics marks

mech	vectors	algebra	analysis	stats
77	82	67	67	81
23	38	36	48	15
63	78	80	70	81
55	72	63	70	68
...	...	...	...	...

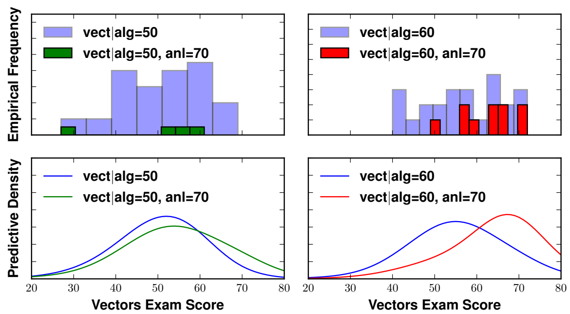
(a) Database of mathematics marks for 88 students, where rows are students and columns are exam scores.

	M	V	G	L	S
M	1.00	0.33	0.23	0.00	0.03
V	0.33	1.00	0.28	0.08	0.02
G	0.23	0.28	1.00	0.43	0.36
L	0.00	0.08	0.43	1.00	0.26
S	0.02	0.02	0.36	0.26	1.00

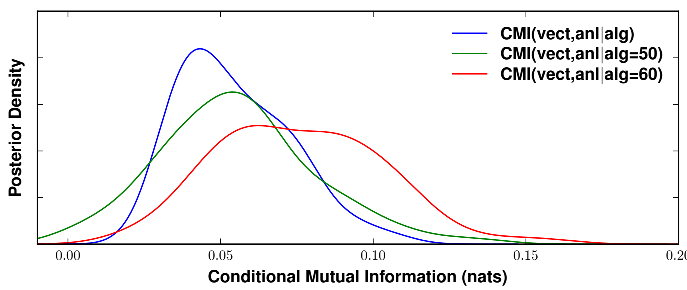
(b) Partial correlation matrix; red entries indicate statistically significant conditional independences.



(c) Undirected (Gaussian) graphical model implied by the partial correlation matrix.



(d) Histograms from the raw dataset (top); and predictive distributions from CrossCat (bottom).



(e) Posterior distribution of CMI(`vectors`, `analysis`) given various conditions of `algebra` show context-specific dependence.

Figure 10: Using posterior CMI distributions to discover *context-specific* predictive relationships in the mathematics marks dataset [20, 34, 6] which are missed by partial correlations. **(a)** The database contains scores of 88 students on five mathematics exams: `mechanics`, `vectors`, `algebra`, `analysis`, and `statistics`. **(b)** Modeling the variables as jointly Gaussian and computing the partial correlation matrix indicates that (`mechanics`, `vectors`) are together conditionally independent of (`analysis`, `statistics`), given `algebra`. **(c)** A Gaussian graphical model which expresses the conditional independences relationships is formed by removing edges whose incident nodes have statistically-significant partial correlations of zero. The graph suggests that when predicting the `vectors` score for a student whose `algebra` score is known, further conditioning on the `analysis` score provides no additional information. We will critique this finding, by showing that the predictive strength of `analysis` on `vectors` given `algebra` varies, depending on the conditioning value of `algebra`. **(d)** The left panel shows that when `algebra` $= 50$, conditioning on `analysis` $= 70$ appears to have little effect on the prediction for `vectors`. The right panel shows that when `algebra` $= 60$, however, conditioning on `analysis` $= 70$ results in a sizeable shift of the posterior mean of `vectors` from 52 to just under 70. This shift is consistent with the top right histogram, where knowing that `analysis` $= 70$ eliminates all the `vectors` scores in the heavy left tail. **(e)** We formalize this "context-specific" dependence by computing the distribution of the CMI of `vectors` and `analysis` under two conditions: `algebra` $= 50$ (green curve), and `algebra` $= 60$ (red curve). The red curve places great probability on higher values of mutual information than the green curve, which explains the shift in predictive density from (d). Finally, we observe that the CMI is weakest when *marginalizing* over all values of `algebra` (blue curve), which explains why the partial correlation of `vectors` and `analysis`, which only considers marginal relationships, is near zero.