entropy

MDPI

Article

Discovering Potential Correlations via Hypercontractivity †

Hyeji Kim [1,2,*], Weihao Gao [1,2], Sreeram Kannan [3], Sewoong Oh [1,4] and Pramod Viswanath [1,2]

1 Coordinated Science Lab, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA; wgao9@illinois.edu (W.G.); swoh@illinois.edu (S.O.); pramodv@illinois.edu (P.V.)
2 Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
3 Department of Electrical Engineering, University of Washington, Seattle, WA 98195, USA; ksreeram@uw.edu
4 Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
* Correspondence: hyejikim@illinois.edu; Tel.: +1-650-644-7087
† This paper is an extended version of our paper given at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.

Received: 11 September 2017; Accepted: 30 October 2017; Published: 2 November 2017

Abstract: Discovering a correlation from one variable to another variable is of fundamental scientific and practical interest. While existing correlation measures are suitable for discovering *average* correlation, they fail to discover hidden or *potential* correlations. To bridge this gap, (i) we postulate a set of natural axioms that we expect a measure of potential correlation to satisfy; (ii) we show that the *rate* of information bottleneck, i.e., the *hypercontractivity* coefficient, satisfies all the proposed axioms; (iii) we provide a novel estimator to estimate the hypercontractivity coefficient from samples; and (iv) we provide numerical experiments demonstrating that this proposed estimator discovers potential correlations among various indicators of WHO datasets, is robust in discovering gene interactions from gene expression time series data, and is statistically more powerful than the estimators for other correlation measures in binary hypothesis testing of canonical examples of potential correlations.

Keywords: correlation analysis; potential correlation; information bottleneck; hypercontractivity

1. Introduction

Measuring the strength of an association between two random variables is a fundamental topic of broad scientific interest. Pearson's correlation coefficient [1] dates from over a century ago and has been generalized seven decades ago as maximal correlation (mCor) to handle nonlinear dependencies [2–4]. Novel correlation measures to identify different kinds of associations continue to be proposed in the literature; these include maximal information coefficient (MIC) [5] and distance correlation (dCor) [6]. Despite the differences, a common theme of measurement of the empirical *average* dependence unites the different dependence measures. Alternatively, these are *factual* measures of dependence and their relevance is restricted when we seek a *potential* dependence of one random variable on another. For instance, consider a hypothetical city with very few smokers. A standard measure of correlation on the historical data in this town on smoking and lung cancer will fail to discover the fact that smoking causes cancer, since the average correlation is very small. On the other hand, clearly, there is a potential correlation between smoking and lung cancer; indeed applications of this nature abound in several scenarios in modern data science, including a recent one on genetic pathway discovery [7].

Discovery of a potential correlation naturally leads one to ask for a measure of potential correlation that is statistically well-founded and addresses practical needs. Such is the focus of

this work, where our proposed measure of potential correlation is based on a novel interpretation of the *Information Bottleneck* (IB) principle [8]. The IB principle has been used to address one of the fundamental tasks in supervised learning: given samples $\{X_i, Y_i\}_{i=1}^n$, how do we find a *compact* summary of a variable X that is most *informative* in explaining another variable Y. The output of the IB principle is a compact summary of X that is most relevant to Y and has a wide range of applications [9,10].

We use this IB principle to create a measure of correlation based on the following intuition: if X is (potentially) correlated with Y, then a relatively compact summary of X can still be very informative about Y. In other words, the maximal ratio of how informative a summary can be in explaining Y to how compact a summary is with respect to X is, conceptually speaking, an indicator of potential correlation from X to Y. Quantifying the compactness by $I(U;X)$ and the information by $I(U;Y)$ we consider the *rate of information bottleneck* as a measure of potential correlation:

$$s(X;Y) \equiv \sup_{U-X-Y} \frac{I(U;Y)}{I(U;X)}, \tag{1}$$

where $U - X - Y$ forms a Markov chain and the supremum is over all summaries U of X. This intuition is made precise in Section 2, where we formally define a natural notion of potential correlation (Axiom 6), and show that the rate of information bottleneck $s(X;Y)$ captures this potential correlation (Theorem 1) while other standard measures of correlation fail (Theorem 2).

This ratio has only recently been identified as the *hypercontractivity* coefficient [11]. Hypercontractivity has a distinguished and central role in a large number of technical arenas including quantum physics [12,13], theoretical computer science [14,15], mathematics [16–18] and probability theory [19,20]. In this paper, we provide a novel interpretation to the hypercontractivity coefficient as a measure of potential correlation by demonstrating that it satisfies a natural set of axioms such a measure is expected to obey.

For practical use in discovering correlations, the standard correlation coefficients are equipped with corresponding natural sample-based estimators. However, for hypercontractivity coefficient, estimating it from samples is widely acknowledged to be challenging, especially for continuous random variables [21–23]. There is no existing algorithm to estimate the hypercontractivity coefficient in general [21], and there is no existing algorithm for solving IB from samples either [22,23]. We provide a novel estimator of the hypercontractivity coefficient—the first of its kind—by bringing together the recent theoretical discoveries in [11,24] of an alternate definition of hypercontractivity coefficient as ratio of Kullback–Leibler divergences defined in (10), and recent advances in joint optimization (the max step in Equation (1)) and estimating information measures from samples using importance sampling [25].

Our main contributions are the following:

- We postulate a set of natural axioms that a measure of potential correlation from X to Y should satisfy (Section 2.1).
- We show that $\sqrt{s(X;Y)}$, our proposed measure of potential correlation, satisfies all the axioms we postulate (Section 2.2). In comparison, we prove that existing standard measures of correlation not only fail to satisfy the proposed axioms, but also fail to capture canonical examples of potential correlations captured by $\sqrt{s(X;Y)}$ (Section 2.3). Another natural candidate is mutual information, but it is not clear how to interpret the value of mutual information as it is unnormalized, unlike all other measures of correlation which are between zero and one.
- Computation of the hypercontractivity coefficient from samples is known to be a challenging open problem. We in troduce a novel estimator to compute hypercontractivity coefficient from i.i.d. samples in a statistically consistent manner for continuous random variables, using ideas from importance sampling and kernel density estimation (Section 3).
- In a series of synthetic experiments, we show empirically that our estimator for the hypercontractivity coefficient is statistically more powerful in discovering a potential correlation than existing correlation

estimators; a larger power means a larger successful detection rate for a fixed false alarm rate (Section 4.1).

- We show applications of our estimator of hypercontractivity coefficient in two important datasets: In Section 4.2, we demonstrate that it discovers hidden potential correlations among various national indicators in WHO datasets, including how aid is potentially correlated with the income growth. In Section 4.3, we consider the following gene pathway recovery problem: we are given samples of four gene expressions time series. Assuming we know that gene A causes B, that B causes C, and that C causes D, the problem is to discover that these causations occur in the sequential order: A to B, and then B to C, and then C to D. We show empirically that the estimator of the hypercontractivity coefficient recovers this order accurately from a vastly smaller number of samples compared to other state-of-the art causal influence estimators.

2. Axiomatic Approach to Measure Potential Correlations

We propose a set of axioms that we expect a measure of potential correlation to satisfy. We then show that hypercontractivity coefficient, first introduced in [19], satisfies all the proposed axioms, hence propose hypercontractivity coefficient as a measure of potential correlation. We also show that other standard correlation coefficients and mutual information, on the other hand, violate the proposed axioms.

2.1. Axioms for Potential Correlation

We postulate that a *measure of potential correlation* $\rho^* : \mathcal{X} \times \mathcal{Y} \to [0,1]$ between two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ should satisfy:

1. $\rho^*(X,Y)$ is defined for any pair of non-constant random variables X and Y.
2. $0 \leq \rho^*(X,Y) \leq 1$.
3. $\rho^*(X,Y) = 0$ iff X and Y are statistically independent.
4. For bijective Borel-measurable functions $f, g : \mathbb{R} \to \mathbb{R}$, $\rho^*(X,Y) = \rho^*(f(X), g(Y))$.
5. If $(X,Y) \sim \mathcal{N}(\mu, \Sigma)$, then $\rho^*(X,Y) = |\rho|$, where ρ is the Pearson correlation coefficient.
6. $\rho^*(X,Y) = 1$ if there exists a subset $\mathcal{X}_r \subseteq \mathcal{X}$ such that for a pair of continuous random variables $(X,Y) \in \mathcal{X}_r \times \mathcal{Y}$, $Y = f(X)$ for a Borel-measurable and non-constant continuous function f.

Axioms 1–5 are identical to a subset of the celebrated axioms of Rényi in [4], which ensure that the measure is properly normalized and invariant under bijective transformations, and recovers the Pearson correlation for jointly Gaussian random variables. Rényi's original axioms for a *measure of correlation* in [4] included Axioms 1–5 and also that the measure ρ^* of correlation should satisfy

6′. $\rho^*(X,Y) = 1$ if for Borel-measurable functions f or g, $Y = f(X)$ or $X = g(Y)$.
7′. $\rho^*(X;Y) = \rho^*(Y;X)$.

The Pearson correlation violates a subset (3, 4, and 6′) of Rényi's axioms. Together with recent empirical successes in multimodal deep learning (e.g., [26–28]), Rényi's axiomatic approach has been a major justification of Hirschfeld–Gebelein–Rényi (HGR) maximal correlation coefficient defined as $\mathrm{mCor}(X,Y) := \sup_{f,g} \mathsf{E}[f(X)g(Y)]$, which satisfies all Rényi's axioms [2]. Here, the supremum is over all measurable functions with $\mathsf{E}[f(X)] = \mathsf{E}[g(Y)] = 0$ and $\mathsf{E}[f^2(X)] = \mathsf{E}[g^2(Y)] = 1$. However, maximal correlation is not the only measure satisfying all of Rényi's axioms, as we show in the following.

Proposition 1. *For any function $F : [0,1] \times [0,1] \to [0,1]$ satisfying $F(x,y) = F(y,x)$, $F(x,x) = x$, and $F(x,y) = 0$ only if $xy = 0$, the symmetrized $F(\sqrt{s(X;Y)}, \sqrt{s(Y;X)})$ satisfies all Rényi's axioms.*

This follows from the fact that the hypercontractivity coefficient $\sqrt{s(X;Y)}$ satisfies all but the symmetry in Axiom 7′ (Theorem 1), and it follows that a symmetrized version satisfies all axioms, e.g., $(1/2)(\sqrt{s(X;Y)} + \sqrt{s(Y;X)})$ and $(s(X;Y)s(Y;X))^{1/4}$. A formal proof is provided in Section 5.1.

From the original Rényi's axioms, for a potential correlation measure, we remove Axiom 7′ that ensures symmetry, as directionality is fundamental in measuring the potential correlation from X to Y. We further replace Axiom 6′ by Axiom 6, as a variable X has a full potential to be correlated with Y if there exists a domain $\mathcal{X}_r$ such that X and Y are deterministically dependent and non-degenerate (i.e., not a constant function), as illustrated in Figure 1 for a linear function and a quadratic function.

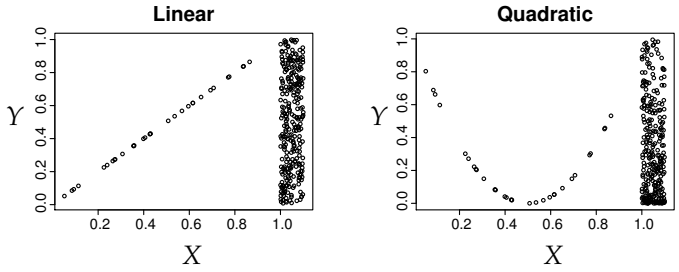

Figure 1. A measure of potential correlation should capture the rare correlation in $X \in [0, 1]$ in these examples which satisfy Axiom 6 for a linear and a quadratic function, respectively.

2.2. The Hypercontractivity Coefficient Satisfies All Axioms

We show that the hypercontractivity coefficient defined in Equation (1) satisfies all Axioms 1–6. Intuitively, $s(X;Y)$ measures how much potential correlation X has with Y. For example, if X and Y are independent, then $s(X;Y) = 0$ as X has no correlation with Y (Axiom 3). By data processing inequality, it follows that it is a measure between zero and one (Axiom 2) and also invariant under bijective transformations (Axiom 4). For jointly Gaussian variables X and Y with the Pearson correlation ρ, we can show that $s(X;Y) = s(Y;X) = \rho^2$. Hence, the squared-root of $s(X;Y)$ satisfies Axiom 5. In fact, $\sqrt{s(X;Y)}$ satisfies all desired axioms for potential correlation, and we make this precise in the following theorem whose proof is provided in Section 5.2.

Theorem 1. *Hypercontractivity coefficient* $\sqrt{s(X;Y)}$ *satisfies Axioms 1–6.*

In particular, the hypercontractivity coefficient satisfies Axiom 6 for potential correlation, unlike other measures of correlation (see Theorem 2 for examples). If there is a potential for X in a possibly rare regime in $\mathcal{X}$ to be fully correlated with Y such that $Y = f(X)$, then the hypercontractivity coefficient is maximum: $s(X;Y) = 1$. In the following section, we show that existing correlation measures, on the other hand, violate the proposed axioms.

2.3. Standard Correlation Coefficients Violate the Axioms

We analyze existing measures of correlations under the scenario with potential correlation (Axiom 6), where we find that none of the existing correlation measures satisfy Axiom 6. Suppose X and Y are independent (i.e., no correlation) in a subset $\mathcal{X}_d$ of the domain $\mathcal{X}$, and allow X and Y to be arbitrarily correlated in the rest $\mathcal{X}_r$ of the domain, such that $\mathcal{X} = \mathcal{X}_d \cup \mathcal{X}_r$. We further assume that the independent part is dominant and the correlated part is rare; let $\alpha := \mathsf{P}(X \in \mathcal{X}_r)$ and we consider the scenario when α is small. A good measure of potential correlation is expected to capture the correlation in $\mathcal{X}_r$ even if it is rare (i.e., α is small). To make this task more challenging, we assume that the conditional distribution of $Y|\{X \in \mathcal{X}_r\}$ is the same as $Y|\{X \notin \mathcal{X}_r\}$. Figure 1 illustrates sampled points for two examples from such a scenario and more examples are in Figure 3. Our main result is the analysis of HGR maximal correlation (mCor) [2], distance correlation (dCor) [6], maximal information coefficients (MIC) [5], which shows that these measures are vanishing with α even if the dependence in the rare regime is very high. Suppose $Y|(X \in \mathcal{X}_r) = f(X)$, then all three correlation coefficients are vanishing as α gets small. This in particular violates Axiom 6. The reason is that standard correlation coefficients measure the *average correlation* whereas the hypercontractivity coefficient measures the

potential correlation. The experimental comparisons on the power of these measures confirm our analytical predictions in Figure 4 in Section 4. The formal statement is below and the proof is provided in Section 5.3.

Theorem 2. *Consider a pair of continuous random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. Suppose $\mathcal{X}$ is partitioned as $\mathcal{X}_r \cup \mathcal{X}_d = \mathcal{X}$ such that $P_{Y|X}(S|X \in \mathcal{X}_r) = P_{Y|X}(S|X \in \mathcal{X}_d)$ for all $S \subseteq \mathcal{Y}$, and Y is independent of X for $X \in \mathcal{X}_d$. Let $\alpha = P\{X \in \mathcal{X}_r\}$. The HGR maximal correlation coefficient is*

$$\mathrm{mCor}(X, Y) \;=\; \sqrt{\alpha}\, \mathrm{mCor}(X_r, Y)\,, \tag{2}$$

the distance correlation coefficient is

$$\mathrm{dCor}(X, Y) \;=\; \alpha\, \mathrm{dCor}(X_r, Y)\,, \tag{3}$$

the maximal information coefficient is upper bounded by

$$\mathrm{MIC}(X, Y) \;\leq\; \alpha\, \mathrm{MIC}(X_r, Y)\,, \tag{4}$$

where X_r is the random variable X conditioned on the rare domain $X \in \mathcal{X}_r$.

Under the rare/dominant scenario considered in Theorem 2, $s(X;Y) \geq \mathrm{mCor}^2(X;Y)$. It is well known that this inequality holds for any X and Y [19]. In particular, Theorem 3 in [29] shows that hypercontractivity coefficient is a natural extension of the popular HGR maximal correlation coefficient as follows.

Remark 1 (Connection between $s(X;Y)$ and $\mathrm{mCor}(X,Y)$ [29])**.** *The squared HGR maximal correlation is a special case of the hypercontractivity optimization in Equation (10) restricted to searching over a distribution $r(x)$ in a close neighborhood of $p(x)$.*

As $s(X;Y)$ searches over a larger space, it is always larger than or equal to $\mathrm{mCor}^2(X;Y)$. This gives an intuitive justification for using $s(X;Y)$ as a measure of potential influence; we allow search over larger space, but properly normalized by the KL divergence, in a hope to find a potential distribution $r(x)$ that can influence Y significantly. While hypercontractivity coefficient is a natural extension of HGR maximal correlation coefficient, there is an important difference between hypercontractivity coefficient and HGR maximal correlation coefficient (and other correlation measures); hypercontractivity is directional.

Remark 2 (Asymmetry of $s(X;Y)$)**.** *Hypercontractivity coefficient is asymmetric in X and Y while HGR maximal correlation, distance correlation, and MIC are symmetric.*

Under the rare/dominant scenario considered in Theorem 2, the hypercontracitivy coefficient $s(X;Y)$ is large because it measures the potential correlation from X to Y. On the other hand, inverse hypercontractivity coefficient $s(Y;X)$, which measures the potential correlation from Y to X, is small as there is no apparent potential correlation from Y to X. This is made precise in the following proposition, with its proof in Section 5.4.

Proposition 2. *Under the hypotheses of Theorem 2, the hypercontractivity coefficient from Y to X is*

$$s(Y;X) = \alpha\, s(Y;X_r),$$

where X_r is the random variable X conditioned on the rare domain $X \in \mathcal{X}_r$.

2.4. Mutual Information Violates the Axioms

Beside standard correlation measures, another measure widely used to quantify the strength of dependence is mutual information. We can show that mutual information satisfies Axiom 6 if we replace 1 by ∞. However there are two key problems:

- Practically, mutual information is *unnormalized*, i.e., $I(X;Y) \in [0,\infty)$. Hence, it provides no absolute indication of the strength of the dependence.
- Mathematically, we are looking for a quantity that *tensorizes*, i.e., does not change when there are many i.i.d. copies of the same pair of random variables.

Remark 3 (Tensorization property of $s(X;Y)$ [30])**.** *Hypercontractivity coefficient tensorizes, i.e.,*

$$s(X_1,...,X_n;Y_1,..,Y_n) = s(X_1,Y_1), \text{ for i.i.d. } (X_i,Y_i), \ i = 1,\cdots,n.$$

On the other hand, mutual information is additive, i.e.,

$$I(X_1,\cdots,X_n;Y_1,\cdots,Y_n) = nI(X_1;Y_1), \text{ for i.i.d. } (X_i,Y_i), \ i = 1,\cdots,n.$$

Tensorizing quantities capture the strongest relationship among independent copies while additive quantities capture the sum. For instance, mutual information could be large because a small amount of information accumulates over many of the independent components of X and Y (when X and Y are high dimensional) while tensorizing quantities would rule out this scenario, where there is no strong dependence. When the components are not independent, hypercontractivity indeed pools information from different components to find the strongest direction of dependence, which is a desirable property.

One natural way to normalize mutual information is by the log of the cardinality of the input/output alphabets [31]. One can interpret a popular correlation measure MIC as a similar effort for normalizing mutual information and is one of our baselines.

Given that other correlation measures and mutual information do not satisfy our axioms, a natural question to ask is whether hypercontractivity is a unique solution that satisfies all the proposed axioms. In the following, we show that the hypercontractivity coefficient is not the only one satisfying all the proposed axioms—just as HGR correlation is not the only measure satisfying Rényi's original axioms.

2.5. Hypercontractivity Ribbon

We show that a family of measures known as *hypercontractivity ribbon*, which includes hypercontractivity coefficient as a special case, satisfy all the axioms. The hypercontractivity ribbon [19,32] is a class of measures parametrized by $\alpha > 0$ as

$$r_\alpha(X;Y) = \sup_{r(x,y)\neq p(x,y)} \frac{D(r(y)\|p(y))}{D(r(x)\|p(x)) + \alpha D(r(y|x)\|p(y|x))}, \tag{5}$$

where $D(r(x)||p(x))$ denotes the KL divergence of $r(x)$ and $p(x)$. An alternative characterization of hypercontractivity ribbon in terms of mutual information is provided in [24,32];

$$r_\alpha(X;Y) = \sup_{p(u|x,y)} \frac{I(U;Y)}{I(U;X) + \alpha I(U;Y|X)}. \tag{6}$$

from which we can see that hypercontractivity coefficient is a special case of hypercontractivity ribbon [11]:

$$s(X;Y) = \lim_{\alpha\to\infty} r_\alpha(X;Y) = \lim_{\alpha\to\infty} s_\alpha(X;Y).$$

Proposition 3. *The (re-parameterized) hypercontractivity ribbon* $s_\alpha(X;Y) := (\alpha\, r_\alpha(X;Y) - 1)/(\alpha - 1)$, *for* $\alpha > 1$, *satisfies Axioms 1–6.*

Proof. By definition, $s_\alpha(X;Y)$ is defined for any pair of non-constant random variables (Axiom 1) and is between 0 and 1 by data processing inequality (Axiom 2). We can show that $s_\alpha(X;Y)$ satisfies Axioms 3 and 4, in a similar way to show $s(X;Y)$ satisfies Axioms 3 and 4. Also, $s_\alpha(X;Y) = \rho^2$ for a jointly Gaussian X,Y with Pearson correlation ρ [24] (Axiom 5). Finally, $s_\alpha(X;Y)$ satisfies Axiom 6 because $r_\alpha(X;Y)$ is non-increasing in α, which implies that $s_\alpha(X;Y) = r_\alpha(X;Y) = 1$ if $s(X;Y) = 1$. □

Although hypercontracitivy ribbon satisfies all axioms, a few properties of the hypercontractivity coefficient make it more attractive than hypercontractivity ribbon for practical use; hypercontractivity coefficient can be efficiently estimated from samples (see Section 3). Hypercontractivity coefficient is a natural extension of the popular HGR maximal correlation coefficient (Remark 1).

2.6. Multidimensional X and Y

In this section, we discuss potential correlation of multidimensional X and Y. While most of the correlation coefficients, including the hypercontractivity coefficient, are well-defined for multi-dimensional X and Y, the axioms are specific to univariate X and Y. To bridge this gap, we propose replacing Axiom 5, as this is the only axiom specific to univariate random variables.

Axiom 5′. If $(X,Y) \sim \mathcal{N}\left(\mu, \Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}\right)$, then $\rho^*(X,Y) = \|\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\|$, where $\|\cdot\|$ is the spectral norm of a matrix.

This recovers the original Axiom 5 when restricted to univariate X and Y. This naturally generalizes both Rényi's axioms and the proposed potential correlation axioms to multidimensional X and Y.

Proposition 4. *Axiom 5′, together with original Rényi's Axioms 1–4, 6′, and 7′, recovers maximal correlation (mCor) as a measure satisfying all Axioms even in this multi-dimensional case. Axiom 5′, together with our proposed Axioms 1–4, and 6, recovers the hypercontractivity coefficient* $\sqrt{s(X;Y)}$ *as a measure satisfying all axioms.*

The second statement in the proposition follows from the analyses of the hypercontractivity coefficient of Gaussian distributions in [33]. A formal proof is provided in Section 5.7.

2.7. Noisy, Discrete, Noisy and Discrete Potential Correlation

In this section, we consider more general scenarios of potential correlation than the one in Axiom 6. We consider (i) noisy potential correlation where $Y = f(X) + Z$ for a Gaussian noise Z for $(X,Y) \in \mathcal{X}_r \times \mathcal{Y}$, (ii) discrete potential correlation, where $\mathcal{X}_r = \{1, \cdots, k\}$, and (iii) noisy discrete potential correlation—a random corruption model. For these three examples, we obtain a lower bound on $s(X;Y)$.

Example 1. *Suppose for* $(X,Y) \in \mathcal{X}_r \times \mathcal{Y}$, $(X,Y) \sim \mathcal{N}(\mathbf{0}, \Sigma)$ *for*

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Then

$$s(X;Y) \geq \frac{\log \frac{1}{1-\rho^2} + \log \frac{1}{1+\rho^2}}{\log \frac{1}{1-\rho^2} + \frac{H(\alpha)}{\alpha}} \tag{7}$$

Proof is in Section 5.5.

We now consider for discrete (X,Y). We start with the case for which X and Y are perfectly correlated for $(X,Y) \in \mathcal{X}_r \times \mathcal{Y}$.

Example 2. *Suppose that for a pair of discrete random variables $(X,Y) \in \mathcal{X} \times \mathcal{Y}$, there exists a subset $\mathcal{X}_r = \{1,2,\cdots,k\} \subseteq \mathcal{X}$ for which $P\{X \in \mathcal{X}_r\} = \alpha$ and $X|\{X \in \mathcal{X}_r\} \sim \text{Unif}[1:k]$ and $Y = X$ for $X \in \mathcal{X}_r$. Then,*

$$s(X;Y) \geq \frac{\log k}{\log k + \log(1/\alpha)}.$$

The inequality holds by considering $r(x) = \mathbb{I}_{\{X=1\}}$ in (10).

We conjecture this lower bound is indeed tight for $\alpha \leq 0.5$ based on numerical simulations. From this lower-bound, we can see the trade-off between k and α. As $k \to \infty$, the lower bounds approaches to 1. As $\alpha \to 1$, the lower bound approaches to 1. As $\alpha \to 0$, the lower bound approaches to 0. In the following, we consider the case where X and Y are not perfectly correlated in $(\mathcal{X}_r \times \mathcal{Y})$ for discrete (X,Y). In particular, we consider a random corruption model for $(\mathcal{X}_r \times \mathcal{Y})$ and obtain a lower bound on $s(X;Y)$.

Example 3. *Consider a random corruption noise model for $(X,Y) \in \mathcal{X}_r \times \mathcal{Y}$, i.e.,*

$$Y = \begin{cases} X_r & w.p.\ 1 - \frac{k}{k-1}\epsilon, \\ Unif[1:k] & w.p.\ \frac{1}{k-1}\epsilon \end{cases}.$$

Then

$$s(X;Y) \geq \frac{(1-\epsilon)\log k(1-\epsilon) + \epsilon \log k\epsilon/(k-1)}{\log(k/\alpha)} = \frac{\log k - H_2(\epsilon) - \epsilon\log(k-1)}{\log(k/\alpha)}. \tag{8}$$

On the other hand,

$$\text{mCor}^2(X;Y) = \alpha\left(1 - \frac{k}{k-1}\epsilon\right)^2, \ 0 \leq \epsilon \leq \frac{k-1}{k}. \tag{9}$$

Proof is in Section 5.6.

In Figure 2, we show plots of lower bounds on $s(X;Y)$ and $\text{mCor}(X;Y)$ in Examples 1–3; from these figures, we can see that $s(X;Y)$ increases as $\rho \to 1$ and $k \to \infty$, and $s(X;Y)$ is larger than $\text{mCor}(X;Y)$ for $\rho \approx 1$ and large k.

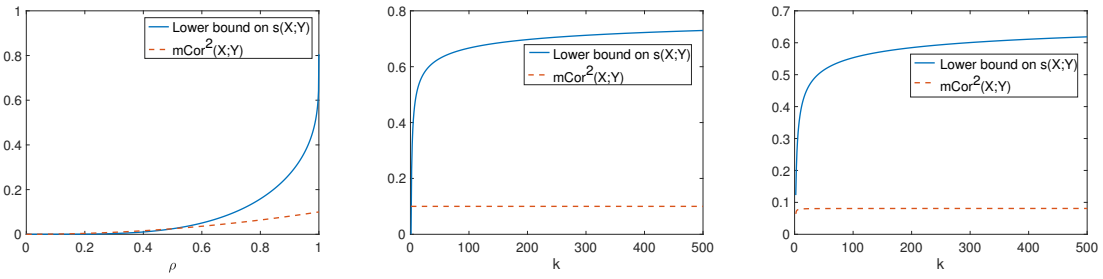

Figure 2. Lower bound on $s(X;Y)$ and $\text{mCor}(X;Y)$ for $\alpha = 0.1$ in (**left**) Example 1 (**middle**) Example 2 (**right**) Example 3 for $\epsilon = 0.1$.

3. Estimator of the Hypercontractivity Coefficient from Samples

In this section, we present an algorithm to compute the hypercontractivity coefficient $s(X;Y)$ from i.i.d. samples $\{X_i, Y_i\}_{i=1}^n$. The computation of the hypercontractivity coefficient from samples is known to be challenging for continuous random variables [22,23], and to the best of our knowledge, there is no known efficient algorithm to compute the hypercontractivity coefficient from samples. Our estimator is the first efficient algorithm to compute the hypercontractivity coefficient, based on the following equivalent definition of the hypercontractivity coefficient, shown recently in [11]:

$$s(X;Y) \equiv \sup_{r_x \neq p_x} \frac{D(r_y||p_y)}{D(r_x||p_x)} . \tag{10}$$

There are two main challenges for computing $s(X;Y)$. The first challenge is – given a marginal distribution r_x and samples from p_{xy}, how do we estimate the KL divergences $D(r_y||p_y)$ and $D(r_x||p_x)$. The second challenge is the optimization over the infinite dimensional simplex. We need to combine estimation and optimization together in order to compute $s(X;Y)$. Our approach is to combine ideas from traditional kernel density estimates and from importance sampling. Let $w_i = r_x(X_i)/p_x(X_i)$ be the *likelihood ratio* evaluated at sample i. We propose the estimation and optimization be solved jointly as follows:

Estimation: To estimate KL divergence $D(r_x||p_x)$, notice that

$$D(r_x||p_x) = \mathbb{E}_{X\sim p_x}\left[\frac{r_x(X)}{p_x(X)} \log \frac{r_x(X)}{p_x(X)}\right].$$

Using empirical average to replace the expectation over p_x, we propose

$$\widehat{D}(r_x||p_x) = \frac{1}{n}\sum_{i=1}^n \frac{r_x(X_i)}{p_x(X_i)} \log \frac{r_x(X_i)}{p_x(X_i)} = \frac{1}{n}\sum_{i=1}^n w_i \log w_i .$$

For $D(r_y||p_y)$, we follow the similar idea, but the challenge is in computing $v_j = r_y(Y_j)/p_y(Y_j)$. To do this, notice that $r_{xy} = r_x p_{y|x}$, so

$$r_y(Y_j) = \mathbb{E}_{X\sim r_x}\left[p_{y|x}(Y_j|X)\right] = \mathbb{E}_{X\sim p_x}\left[p_{y|x}(Y_j|X)\frac{r_x(X)}{p_x(X)}\right] .$$

Replacing the expectation by empirical average again, we get the following estimator of v_j:

$$\widehat{v}_j = \frac{1}{n}\sum_{i=1}^n \frac{p_{y|x}(Y_j|X_i)}{p_y(Y_j)}\frac{r_x(X_i)}{p_x(X_i)} = \frac{1}{n}\sum_{i=1}^n \underbrace{\frac{p_{xy}(X_i,Y_j)}{p_x(X_i)p_y(Y_j)}}_{A_{ji}} w_i .$$

We can write this expression in matrix form as $\widehat{\mathbf{v}} = \mathbf{A}^T\mathbf{w}$. We use a kernel density estimator from [34] to estimate the matrix $\mathbf{A}$, but our approach is compatible with any density estimator of choice.

Optimization: Given the estimators of the KL divergences, we are able to convert the problem of computing $s(X;Y)$ into an optimization problem over the vector $\mathbf{w}$. Here a constraint of $(1/n)\sum_{i=1}^n w_i = 1$

is needed to satisfy $\mathbb{E}_{p_x}[r_x/p_x] = 1$. To improve numerical stability, we use $\log s(X;Y)$ as the objective function. Then the optimization problem has the following form:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \log\Big((\mathbf{w}^T \mathbf{A} \log(\mathbf{A}^T \mathbf{w}) \Big) - \log\Big(\mathbf{w}^T \log \mathbf{w} \Big) \\ \text{subject to} \quad & \frac{1}{n}\sum_{i=1}^{n} w_i = 1 \\ & w_i \geq 0, \forall\, i \end{aligned}$$

where $\mathbf{w}^T \log \mathbf{w} = \sum_{i=1}^{n} w_i \log w_i$ for short. Although this problem is not convex, we apply gradient descent to maximize the objective. In practice, we initialize $w_i = 1 + \mathcal{N}(0, \sigma^2)$ for $\sigma^2 = 0.01$. Hence, the initial r_x is perturbed mildly from p_x. Although we are not guaranteed to achieve the global maximum, we consistently observe in extensive numerical experiments that we have 50–60% probability of achieving the same maximum value, which we believed to be the global maximum. A theoretical analysis of the landscape of local and global optima and their regions of attraction with respect to gradient descent is an interesting and challenging open question, outside the scope of this paper.

Consistency of Estimation

While a theoretical understanding of the performance of gradient descent on the optimization step (where the number of samples is fixed) above is technically very challenging, we can study the performance of the solution as the number of samples increases. In particular we show below (under suitable simplifying assumptions to get to the essence of the proof) that the optimal solution to the finite sample optimization problem is *consistent*. Suppose that $\mathcal{X}$ is discrete. Further we restrict the optimization over a quantized and bounded set T_Δ, where $\mathbf{w} \in T_\Delta$ is quantized by a gap Δ and satisfies: (1) $C_1 \leq w_i \leq C_2$ for all i; (2) $(1/n)\sum_{i=1}^{n} w_i \log w_i > C_0$. We further assume that we have access of $\mathbf{A} = P_{xy}(X_i, Y_j)/P_x(X_i)P_y(Y_j)$. Define $\widehat{s}_\Delta(X;Y) = \max_{\mathbf{w}\in T_\Delta} \mathbf{w}^T \mathbf{A} \log(\mathbf{A}^T \mathbf{w})/\mathbf{w}^T \log \mathbf{w}$, then with two further simplifying conditions on the joint distribution (formally stated in Section 5.8), we can prove consistency of our estimation procedure:

Theorem 3. *As n goes to infinity, $\widehat{s}_\Delta(X;Y)$ converges to $s(X;Y)$ up to a resolution of quantization in probability, i.e., for any $\varepsilon > 0$, $\Delta > 0$ and $s(\Delta) = O(\Delta)$, we have*

$$\lim_{n\to\infty} \mathcal{P}\left(\left| \widehat{s}_\Delta(X;Y) - s(X;Y) \right| > \varepsilon + s(\Delta) \right) = 0\,. \tag{11}$$

4. Experimental Results

We present experimental results on synthetic and real datasets showing that the hypercontractivity coefficient (a) is more powerful in detecting potential correlation compared to existing measures; (b) discovers hidden potential correlations among various national indicators in WHO datasets; and (c) is more robust in discovering pathways of gene interactions from gene expression time series data.

4.1. Synthetic Data: Power Test on Potential Correlation

As our estimator (and the measure itself) involves a maximization, it is possible that we are sensitive to outliers and may capture spurious noise. Via a series of experiments we show that the hypercontractivity coefficient and our estimator are capturing the true potential correlation. As shown in Figure 3, we generate pairs of datasets—one where X and Y are independent and one where there is a potential correlation as per our scenario. We run experiment with eight types of functional associations, following the examples from [5,35,36]. For the correlated datasets, out of n samples $\{(x_i, y_i)\}_{i=1}^{n}$, αn rare but correlated samples are in $\mathcal{X} = [0,1]$ and $(1-\alpha)n$ dominant

but independent samples are in $\mathcal{X} \in [1, 1.1]$. The rare but correlated samples are generated as $x_i \sim \text{Unif}[0,1], y_i \sim f(x_i) + \mathcal{N}(0, \sigma^2)$ for $i \in [1 : \alpha n]$. The dominant samples are generated as $x_i \sim \text{Unif}[1, 1.1], y_i \sim f(\text{Unif}[0,1]) + \mathcal{N}(0, \sigma^2)$ for $i \in [\alpha n + 1, n]$.

Table 1 shows the hypercontractivity coefficient and the other correlation coefficients for correlated and independent datasets shown in Figure 3, along with the chosen value of α and σ^2. Correlation estimates with the largest separation for each row is shown in bold. The hypercontractivity coefficient gives the largest separation between the correlated and the independent dataset for most functional types.

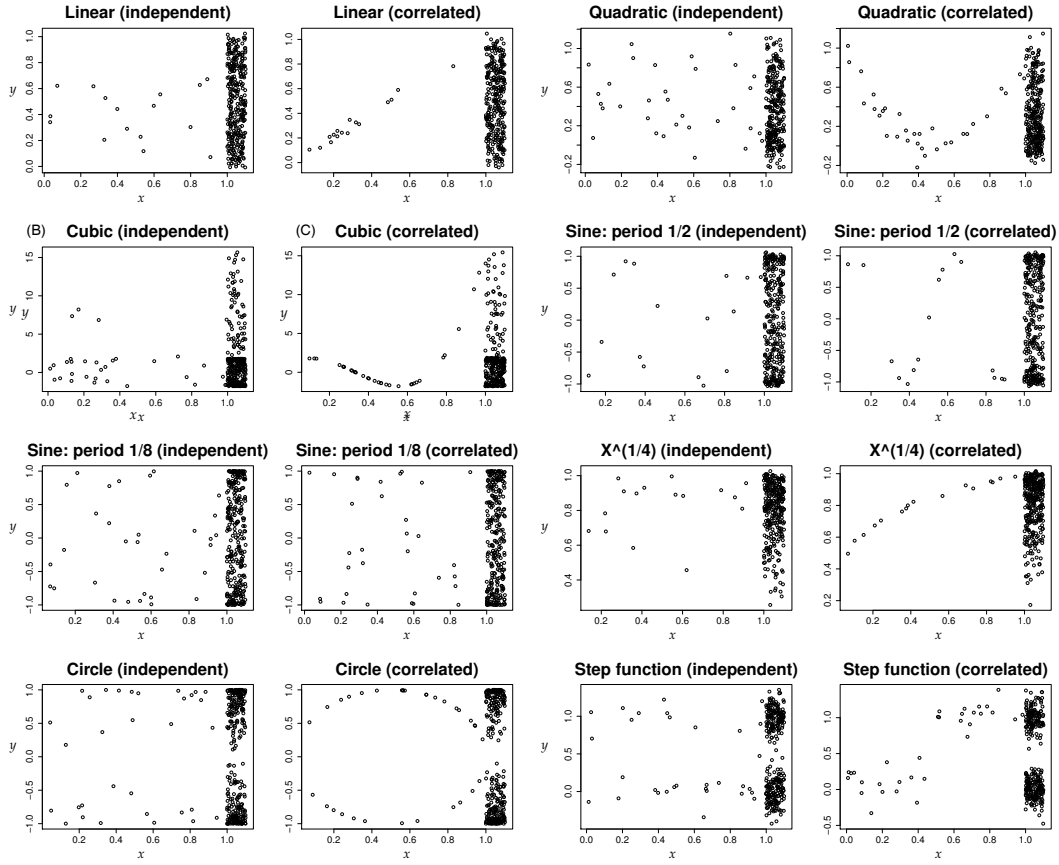

Figure 3. Sample data points for eight functions with/without a potential correlation for $n = 320$.

Table 1. Correlation estimates for independent and correlated samples from Figure 3.

				Cor		dCor		mCor		MIC		HC	
#	Function	α	σ^2	Dep	Indep	Dep	Indep	Dep	Indep	Dep	Indep	Dep	Indep
1	Linear	0.05	0.03	0.03	0.00	0.19	0.11	0.06	0.04	0.21	0.17	**0.18**	**0.08**
2	Quadratic	0.10	0.10	0.00	0.01	0.09	0.10	**0.07**	**0.02**	0.21	0.18	0.08	0.04
3	Cubic	0.10	0.00	0.02	0.00	0.16	0.08	0.09	0.03	**0.26**	**0.17**	0.11	0.04
4	$\sin(4\pi X)$	0.05	0.03	0.00	0.00	0.10	0.06	0.03	0.01	0.20	0.18	**0.10**	**0.04**
5	$\sin(16\pi X)$	0.10	0.00	0.00	0.00	0.07	0.08	**0.03**	**0.03**	0.18	0.22	**0.03**	**0.03**
6	$X^{1/4}$	0.05	0.01	0.01	0.00	0.12	0.07	0.02	0.01	0.20	0.20	**0.12**	**0.04**
7	Circle	0.10	0.00	0.00	0.00	0.09	0.05	0.01	0.03	0.16	0.17	**0.06**	**0.01**
8	Step func.	0.10	0.03	0.00	0.00	0.13	0.07	0.04	0.02	0.20	0.17	**0.11**	**0.04**

A formal statistical approach to test the robustness as well as accuracy is to run *power tests*: testing for the power of the estimator in binary hypothesis tests. To compute the power of each estimator, we compare the false negative rate at a fixed false positive rate of, say, 5%. We generate 500 independent datasets and 500 correlated datasets. We compute the correlation estimates on 500 independent samples, and take the top 5% as a threshold. We compute the correlation estimates on 500 correlated samples. Power is defined as the fraction of correlated datasets for which the correlation estimate is larger than the threshold.

We show empirically that for linear, quadratic, sine with period $1/2$, and the step function, the hypercontractivity coefficient is more powerful as compared to other measures. For a given setting, a larger power means a larger successful detection rate for a fixed false alarm rate. Figure 4 shows the power of correlation estimators as a function of the additive noise level, σ^2, for $\alpha = 0.05$ and $n = 320$. The hypercontractivity coefficient is more powerful than other correlation estimators for most functions. The power of all the estimators are very small for sine (period $1/8$) and circle functions. This is not surprising given that it is very hard to discern the correlated and independent cases even visually, as shown in Figure 3.

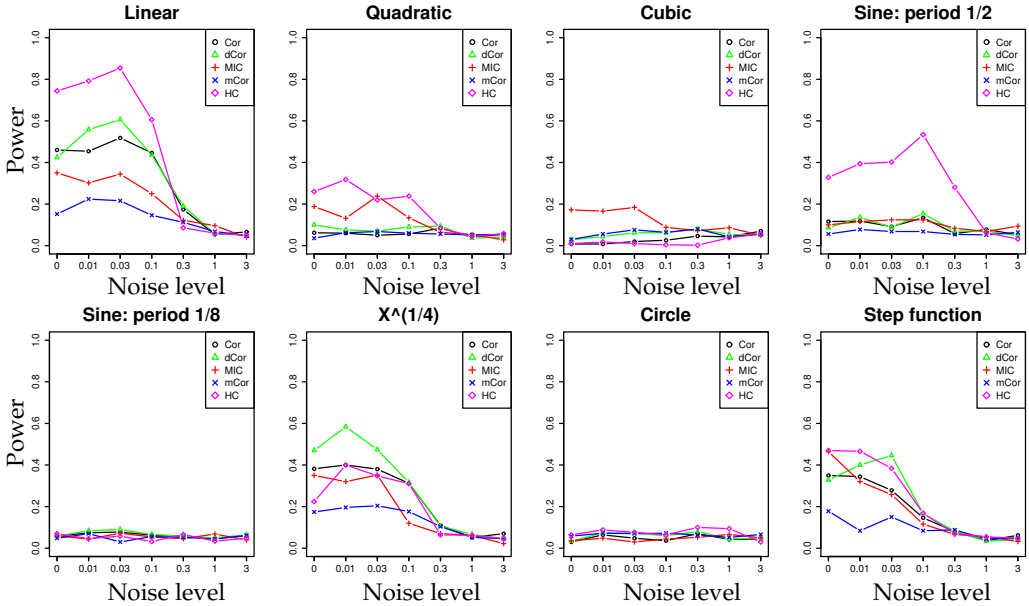

Figure 4. Power vs. noise level for $\alpha = 0.05$, $n = 320$.

Figure 5 plots the power of correlation estimators as a function of noise level for $\alpha = 0.1$ and $n = 320$. As we can see from these figures, hypercontractivity estimator is more powerful than other correlation estimators for most functions. For circle function, the gap between the power of hypercontractivity estimator and the powers of other estimators is significantly large.

On the other hand, hypercontractivity estimator is power deficient for the cubic function. This is because in estimating hypercontractivity coefficient, we estimate $p(y_j|x_i)/p(y_j)$ using the kernel density estimator (KDE), which gives a smooth estimate of $p(y_j|x_i)/p(y_j)$, i.e., for x_i and x_j close to each other, estimated $p(y|x_i)$ and $p(y|x_j)$ are close to each other. Hence, for a correlated dataset for a cubic function, shown in Figure 6C, the estimated $p(y|x)$ does not vary much for x. (Estimated $p(y|x)$ for $x \in [0.8 : 1]$ and $p(y|x)$ for $x \in [1 : 1.1]$ are close to each other). This results in a small hypercontracitivy, which in turn results in a low power in the hypothesis testing. To further analyze this effect, we considered the same dataset but with dominant independent samples appear on the left, as shown in Figure 6E,F, and computed the power of hypercontractivity estimator, shown in Figure 6D. Hypercontractivity estimator is much more powerful than the one for the original dataset. This is because the estimated

$p(y|x)$ for $x \in [0.8, 1]$ is very different from the estimated $p(y|x)$ for $x \in [-0.1, 0]$, which results in a large estimate of hypercontractivity coefficient for the correlated dataset.

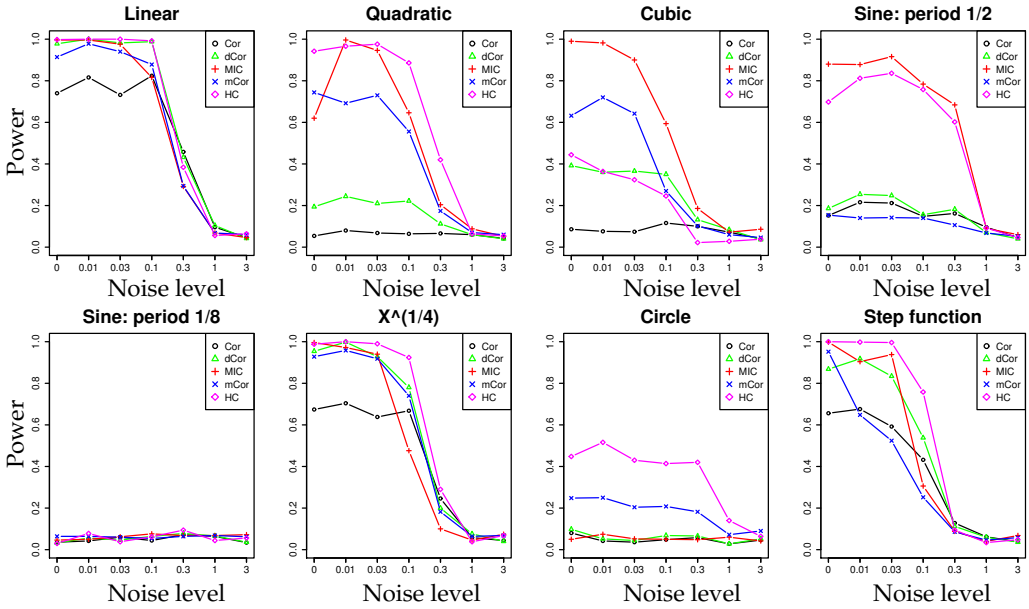

Figure 5. Power vs. noise level for $\alpha = 0.1$, $n = 320$.

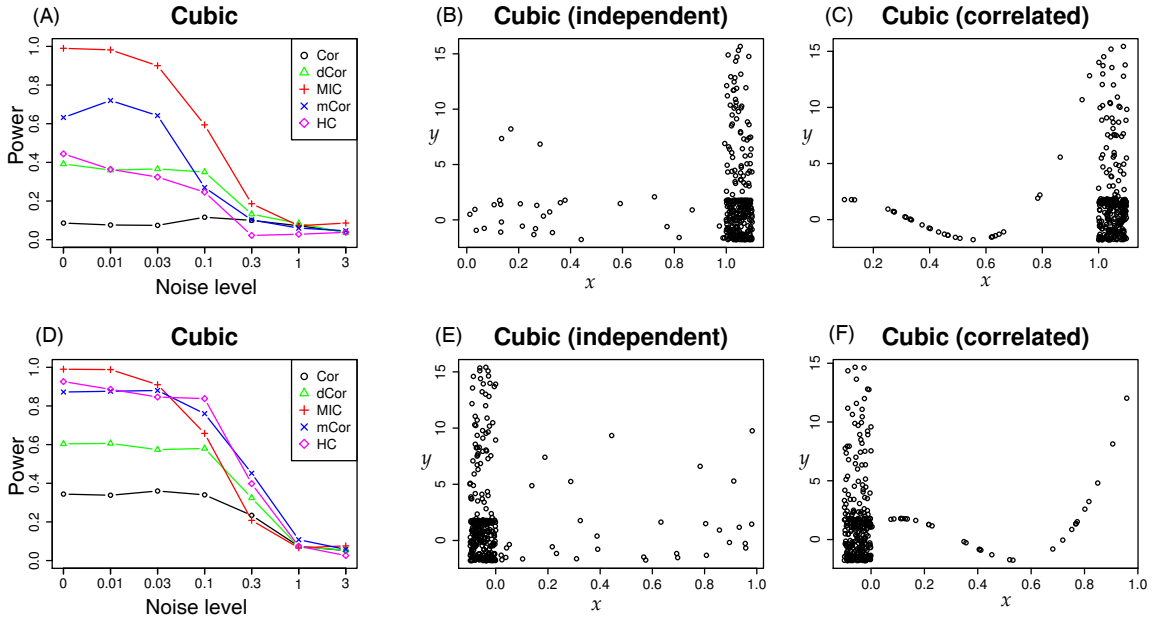

Figure 6. Power vs. noise level for $\alpha = 0.1$ and $n = 320$ (**A**,**D**), corresponding examples of an independent dataset (**B**,**E**) and a correlated dataset (**C**,**F**).

To investigate the dependency of power on α more closely, in Figure 7, we plot the power vs. α or $n = 320$ and $\sigma^2 = 0.1$. Hypercontractivity estimator is more powerful than other estimators for most α, for all functions except for cubic function. For a sine with period $1/8$, due to its high frequency, the powers of all the correlation estimators do not increase as α increases. Figure 8 plots the power vs. sample size n for $\alpha = 0.05$ and $\sigma^2 = 0.1$. For sine with period $1/2$, hypercontractivity estimator is

much more powerful than the other estimators for all sample sizes. We can also see that for sine with period 1/8, powers of all correlation estimators do not increase as sample size increases.

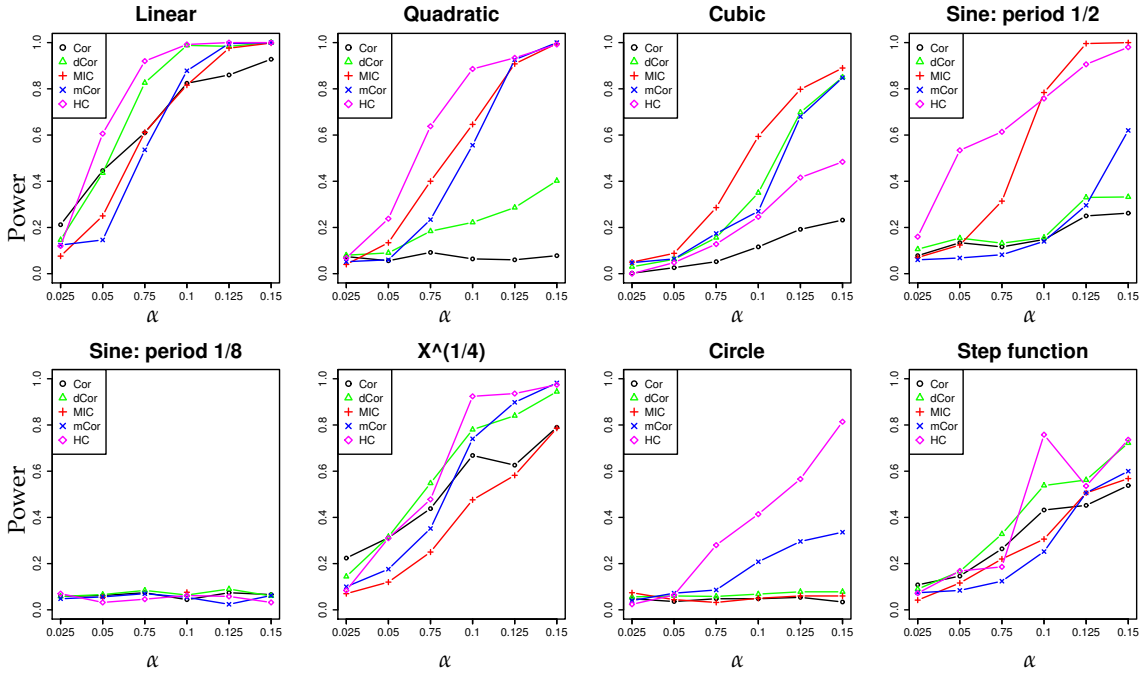

Figure 7. Power vs. α (fraction of correlated samples) for $n = 320$, $\sigma^2 = 0.1$.

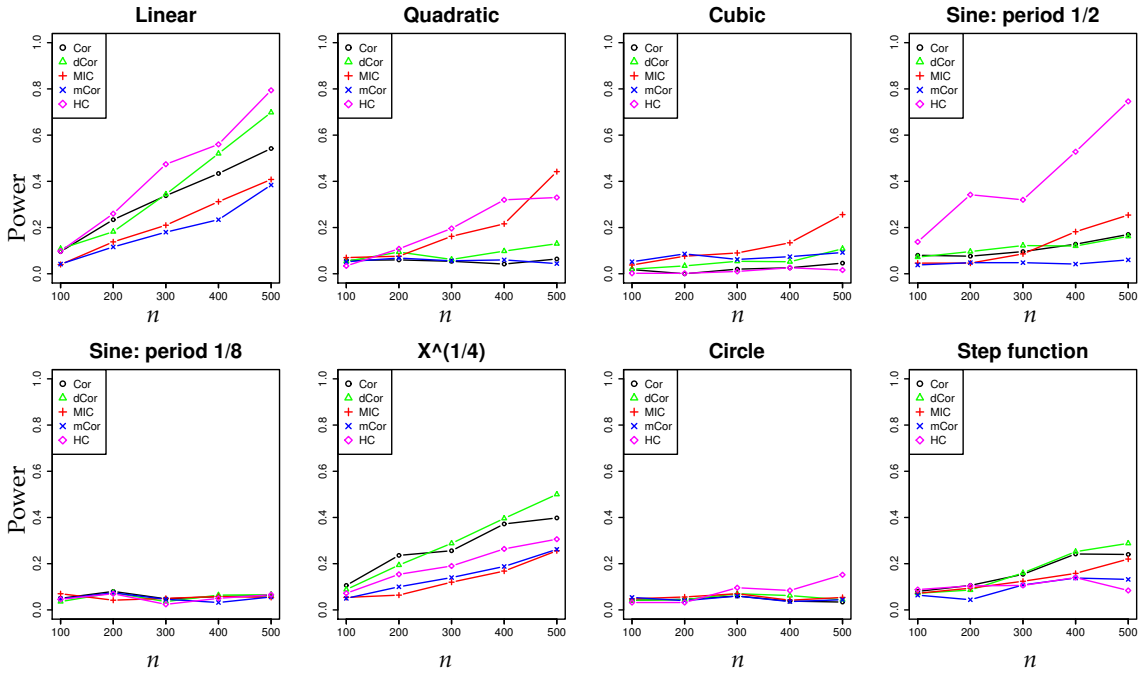

Figure 8. Power vs. n (number of samples) for $\alpha = 0.05$, $\sigma^2 = 0.1$.

4.2. Real Data: Correlation between Indicators of WHO Datasets

We computed the hypercontractivity coefficient, MIC, and Pearson correlation of 1600 pairs of indicators from 202 samples (countries, non i.i.d.) in the World Health Organization (WHO) dataset [5]. Figure 9 illustrates that the hypercontractivity coefficient discovers hidden potential correlation (e.g., in

Figure 9E–H, whereas other measures fail. Scatter plots of Pearson correlation vs. the hypercontractivity coefficient and MIC vs. the hypercontractivity coefficient for all pairs are presented in Figure 9A,D. The samples for pairs of indicators corresponding to B,C,E–J in Figure 9A,D are shown in Figure 9B,C,E–J, respectively. In Figure 9B, it is reasonable to assume that the number of bad teeth per child is uncorrelated with the democracy score. The hypercontractivity coefficient, MIC, and Pearson correlation are all small, as expected. In Figure 9C, the correlation between CO_2 emissions and energy use is clearly visible, and all three correlation estimates are close to one.

However, only the hypercontractivity coefficient discovers the hidden potential correlation in Figure 9E–H. In Figure 9E, the data is a mixture of two types of countries—one with small amount of aid received (less than $\$5 \times 10^8$), and the other with large amount of aid received (larger than $\$5 \times 10^8$). Dominantly many countries (104 out of 146) belong to the first type (small aid), and for those countries, the amount of aid received and the income growth are independent. For the remaining countries with larger aid received, although those are rare, there is a clear correlation between the amount of aid received and the income growth.

Similarly in Figure 9F, there are two types of countries—one with small arms exports (less than $\$2 \times 10^8$) and the other with large arms exports (larger than $\$2 \times 10^8$). Dominantly many countries (71 out of 82) belong to the first type, for which the amount of arms exports and the health expenditure are independent. For the remaining countries that belong to the second type, on the other hand, there is a visible correlation between the arms exports and the health expenditure. This is expected as for those countries that export arms the GDP is positively correlated with both arms exports and health expenditure, whereas for those do not have arms industry, these two will be independent.

In Figure 9G, for dominant number of countries, the number of male deaths from the colon and rectum cancer is small (for 145 out of 169 countries, it is smaller than 2000, while for the remaining countries, it is between 2000 and 50,000), and it is independent of the amount of health expenditure. On the other hand, for the remaining countries with larger number of male deaths from colon and rectum cancer, the two indicators are positively associated. This is expected as both indicators are positively correlated with the population. Only hypercontractivity discovers this hidden potential correlation. MIC and Pearson correlation are small.

In Figure 9H, for dominant number of countries, the number of broadband subscribers is very small and is independent of the private health expenditure; 155 out of 180 countries have broadband subscribers less than 10^6. On the other hand, for the remaining countries, the number of broadband subscribers is positively correlated with the private health expenditure. This is as expected because both indicators are positively correlated with the population. Hypercontractivity is large for this dataset, discovering the hidden correlation, whereas all other correlations all small.

In Figure 9I, most countries do not have large hydroelectricity facilities, and for those countries, energy use and hydroelectricity consumption are independent (41 out of 53 countries have hydroelectricity ≤ 0.25). On the other hand, for the countries which have hydroelectrocity facilities, the amount of total energy use and the amount of hydroelectricity consumption are positively correlated. Hypercontractivity discovers this hidden potential correlation. Unlike in Figure 9G,H for which the fraction of correlated samples was only about 14%, in Figure 9I, the fraction of correlated samples is about 23%. Hence, Pearson correlation is larger compared to Pearson correlation values for Figure 9G,H.

In Figure 9J, there is one country (Luxembourg) with very large amounts of foreign direct investment net inflow and outflow. Due to this outlier, Pearson correlation is close to 1. Hypercontractivity is also close to 1, whereas MIC is small. To analyze the effect of the outlier in correlation measures, in the following, we compute the correlation measures for samples without an outlier.

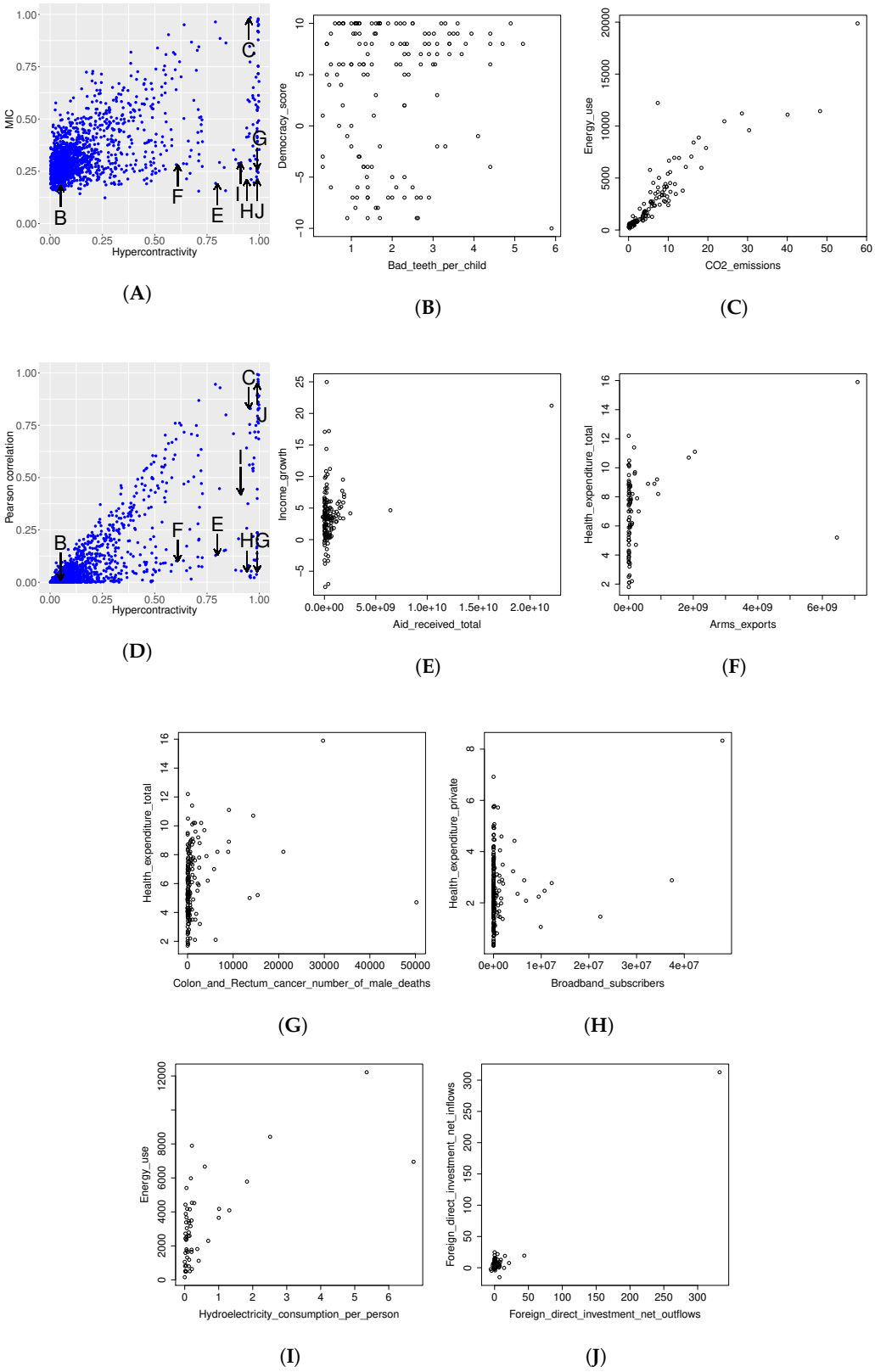

Figure 9. (**A**,**D**) Scatter plot of correlation measures; (**B**) Correlations are small; (**C**) Correlations are large; (**E**–**H**) Only the hypercontractivity coefficient discovers potential correlation; (**I**) Hypercontractivity discovers potential correlation; (**J**) Hypercontractivity and Pearson correlation are large because of an outlier.

4.2.1. How Hypercontractivity Changes as We Remove Outliers

Figures 10–15, on the left, are shown samples from Figure 9E–J respectively. On the middle and on the right are shown all samples but one outlier and all samples but two outliers, respectively. By comparing the hypercontractivity coefficients for the three datasets for each pair of indicators, we can analyze the effect of outliers on hypercontractivity. For a comparison, on the top of each figure, we show the estimated hypercontractivity (HC), MIC, Pearson correlation (Cor), distance correlation (dCor), maximal correlation (mCor), and the hypercontractivity for reversed direction (HCR). In Figures 10 and 11, we can see that hypercontractivity is more sensitive to an outlier than other correlation measures.

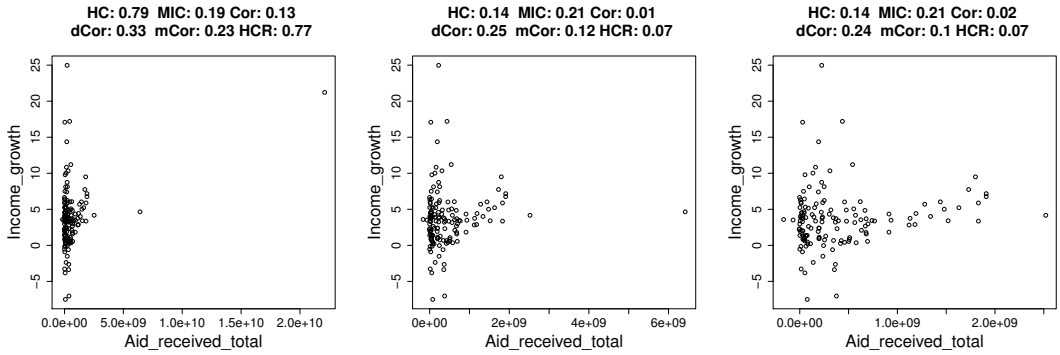

Figure 10. Samples for the pair of indicators shown in Figure 9E from the entire WHO dataset (**left**), without one outlier (**middle**), and without two outliers (**right**).

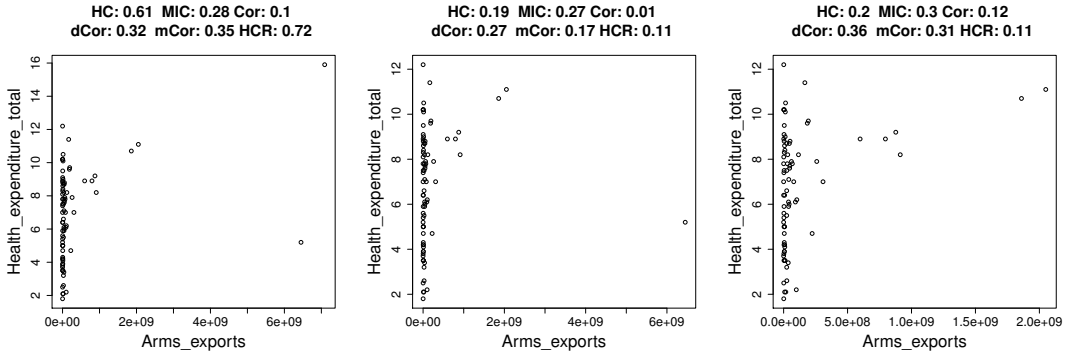

Figure 11. Samples for the pair of indicators shown in Figure 9F from the entire WHO dataset (**left**), without one outlier (**middle**), and without two outliers (**right**).

In Figure 12 (left), the two countries with the largest number of male deaths from the colon and rectum cancer are China and United States. As China is removed from the dataset, in (middle), hypercontractivity remains unchanged. As we also remove United States, in (right), hypercontractivity becomes small, 0.17. This value is still larger than the typical coefficient for two independent indicators ($\approx$0.05), we can see that hypercontractivity is more sensitive to the outlier than other correlation measures.

In Figure 13, the two countries with the largest number of broadband subscribers are United States and China. When we remove United States from the samples, hypercontractivity becomes close to zero, which also shows hypercontractivity is sensitive to the outliers.

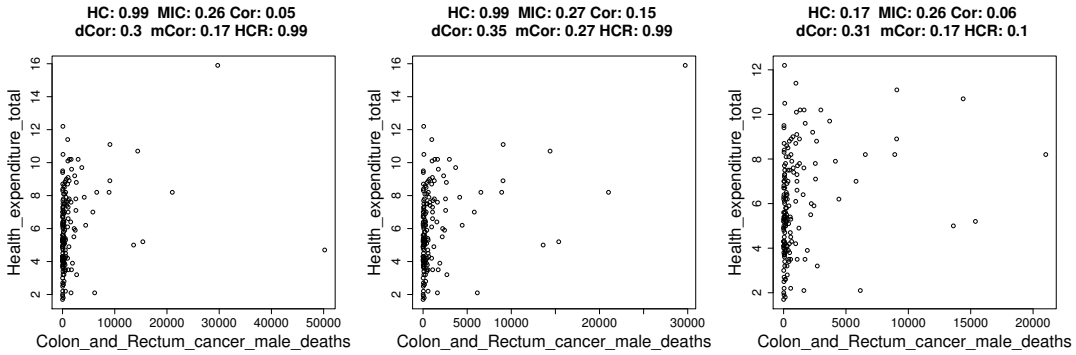

Figure 12. Samples for the pair of indicators shown in Figure 9G from the entire WHO dataset (**left**), without one outlier (**middle**), and without two outliers (**right**).

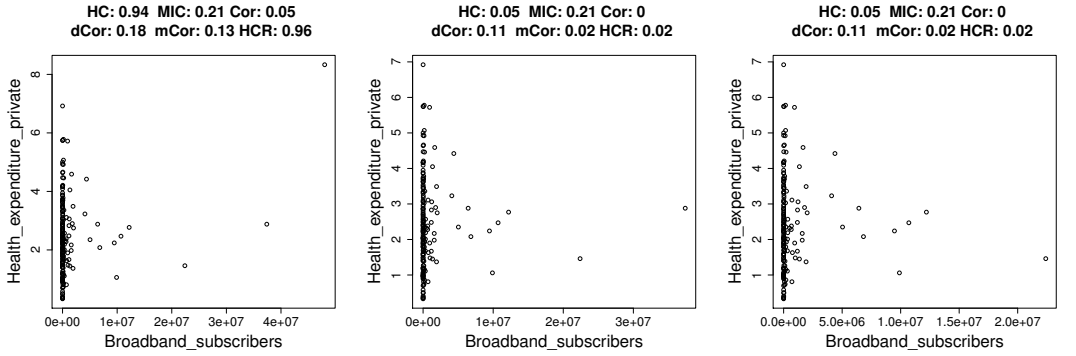

Figure 13. Samples for the pair of indicators shown in Figure 9H from the entire WHO dataset (**left**), without one outlier (**middle**), and without two outliers (**right**).

In Figure 14, hypercontractivity remains large even after we remove outliers. The two countries with the largest amount of hydroelectricity consumption are Norway and Iceland. Even after we remove Norway from the samples, as shown in (middle), hypercontractivity remains large. As we further remove one outlier (Iceland) from the samples, as shown in (right), hypercontractivity becomes 0.49.

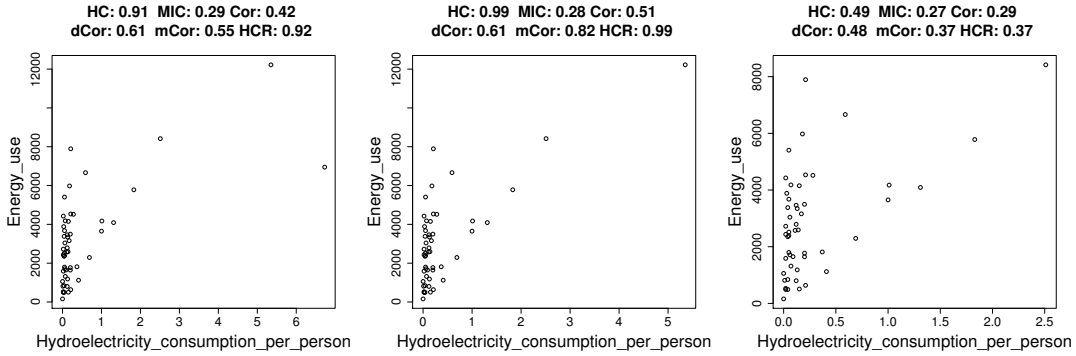

Figure 14. Samples for the pair of indicators shown in Figure 9I from the entire WHO dataset (**left**), without one outlier (**middle**), and without two outliers (**right**).

In Figure 15 (middle), all samples but Luxembourg is shown. We can see that most countries have a very small absolute amount of foreign direct investment net outflows (For 126 out of 157 countries, it is between $[-2, 2]$), and for those countries, the foreign direct investment net outflow is independent

of foreign direct investment net inflows. For the remaining countries, there is a positive association between the outflow and the inflow. Hypercontractivity captures this hidden correlation better than other correlations; hypercontractivity is 0.47, whereas MIC and Pearson correlation are small. If we further remove the rightmost sample, as shown in (right), hypercontractivity becomes small.

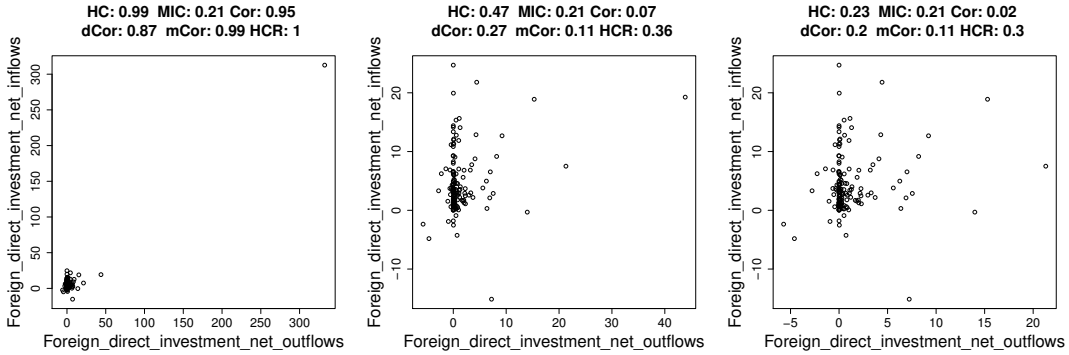

Figure 15. Samples for the pair of indicators shown in Figure 9J from the entire WHO dataset (**left**), without one outlier (**middle**), and without two outliers (**right**).

Whether we should consider a sample in a rare type as a meaningful sample or as an outlier depends on the application. If we use hypercontractivity to discover a pair of measures for which one variable can be potentially correlated with the other, then we would expect to discover that an aid for a country has potential correlation in the income growth. Other measures will fail. It is possible that hypercontractivity might have larger false positive rate, and depending on the application, one might prefer to error on the side of having more positive cases to be screened by further experiments, surveys, or human judgements.

4.2.2. Hypercontractivity Detecting an Outlier

In Figure 16A and B, we show examples of pairs of indicators for which there is one outlier and the remaining samples are independent. In Figure 16C,D, we show the scatterplot of the same pairs of indicators as in Figure 16A,B, respectively, with one outlier removed. As shown in Figure 16A,B, hypercontractivity is close to 1, when there is an outlier. As shown in Figure 16B,D, hypercontractivity is close to 0, when the outlier is removed. This implies that one single outlier can make the hypercontractivity large. We can see similar patterns for other correlation measures, such as Pearson correlation, distance correlation, and maximal correlation from Figure 16, and MIC from Figure 16C,D. Nonetheless, these other correlation measures are not as sensitive to an outlier as hypercontractivity coefficient.

To further study how hypercontractivity estimator is affected by outliers, we ran simulations on synthetic data. We generated three sets of synthetic data shown in Figure 17 and computed hypercontractivity coefficients. In Figure 17 (left), an outlier is located far from the rest of samples, and the estimated hypercontractivity coefficient is 0.99. In Figure 17 (middle), an outlier is located close to the rest of samples, and the estimated hypercontractivity coefficient is 0.04. In Figure 17 (right), X and Y are potentially correlated, and the hypercontractivity estimate is 0.17. As can bee seen from this simulation and experimental results on WHO dataset, our hypercontractivity estimator is sensitive to outliers. If one wants to filter out the effect of outliers, one can combine methods for robust estimation, such as trimming and winsorizing [37–39], along with the hypercontractivity estimator. This is an interesting future research direction.

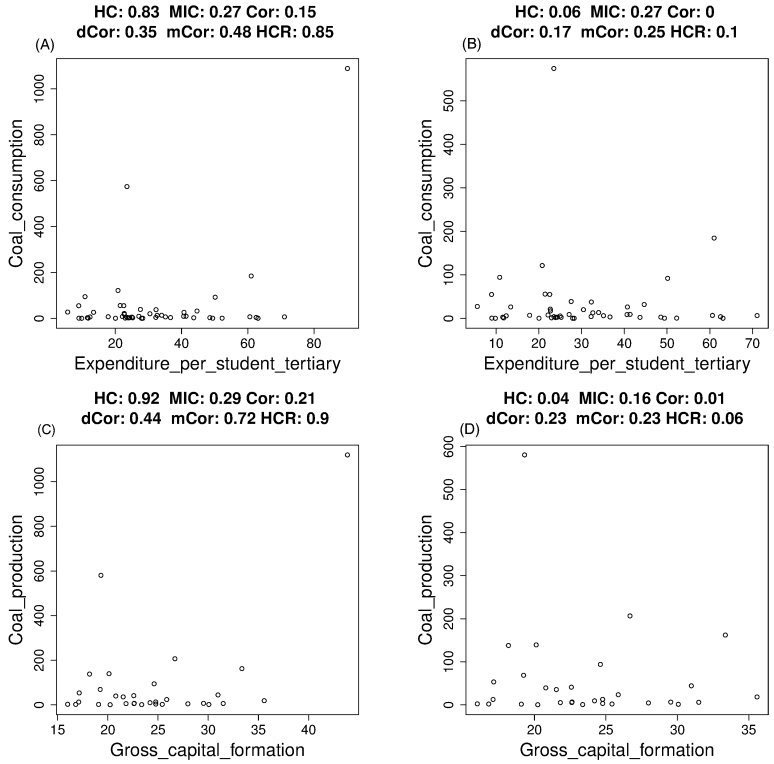

Figure 16. Hypercontractivity and other correlation measures become smaller as we remove an outlier.

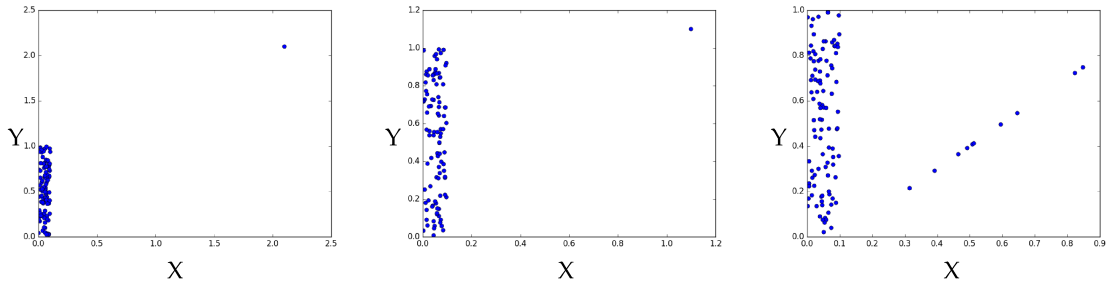

Figure 17. Synthetic data: (**left**) an outlier is located far from other samples; (**middle**) an outlier is located close to the rest of samples; (**right**) potential correlation exists.

4.3. Gene Pathway Recovery From Single Cell Data

We replicate the genetic pathway detection experiment from [7], and show that hypercontractivity correctly discovers the genetic pathways from smaller number of samples. A genetic pathway is a series of genes interacting with each other as a chain. Consider the following setup where four genes whose expression values in a single cell are modeled by random processes X_t, Y_t, Z_t and W_t respectively. These 4 genes interact with each other following a pathway $X_t \to Y_t \to Z_t \to W_t$; it is biologically known that X_t causes Y_t with a negligible delay, and later at time t', $Y_{t'}$ causes $Z_{t'}$, and so on. Our goal is to recover this known gene pathway from sampled datapoints. For a sequence of time points $\{t_i\}_{i=0}^m$, we observe n_i i.i.d. samples $\{X_{t_i}^{(j)}, Y_{t_i}^{(j)}, Z_{t_i}^{(j)}, W_{t_i}^{(j)}\}_{j=1}^{n_i}$ generated from the random process $P(X_{t_i}, Y_{t_i}, Z_{t_i}, W_{t_i})$. We use the real data obtained by the single-cell mass flow cytometry technique [7].

Given these samples from time series, the goal of [7] is to recover the direction of the interaction along the known pathway using correlation measures as follows, where they proposed a new measure called DREMI. The DREMI correlation measure is evaluated on each pairs on the pathway,

$\tau(X_{t_i}, Y_{t_i})$, $\tau(Y_{t_i}, Z_{t_i})$ and $\tau(Z_{t_i}, W_{t_i})$, at each time points t_i. It is declared that a genetic pathway is correctly recovered if the peak of correlation follows the expected trend: $\arg\max_{t_i} \tau(X_{t_i}, Y_{t_i}) \leq \arg\max_{t_i} \tau(Y_{t_i}, Z_{t_i}) \leq \arg\max_{t_i} \tau(Z_{t_i}, W_{t_i})$. In [25], the same experiment has been done with τ evaluated by UMI and CMI estimators. In this paper, we evaluate τ using our proposed estimator of hypercontractivity.

The Figure 18 shows the scatter plots pCD3ζ-pSLP76-pERK-pS6 chain at different time points after TCR activation. The data comes from CD4+ naïve T lymphocytes from B6 mice with CD3, CD28, and CD4 cross-linking. Each row represents a pair of data in the chain, and each column stands for a time point after TCR activation. Estimate of hypercontractivity is shown below the scatter plot for each pair of data and each time point and we highlight the time point where each pair of data is maximally correlated. We can see that the peak of the correlation of pCD3ζ-pSLP76, pSLP76-pERK and pERK-pS6 appears at 0.5 min, 1 min and 2 min respectively, hence the pathway is correctly identified.

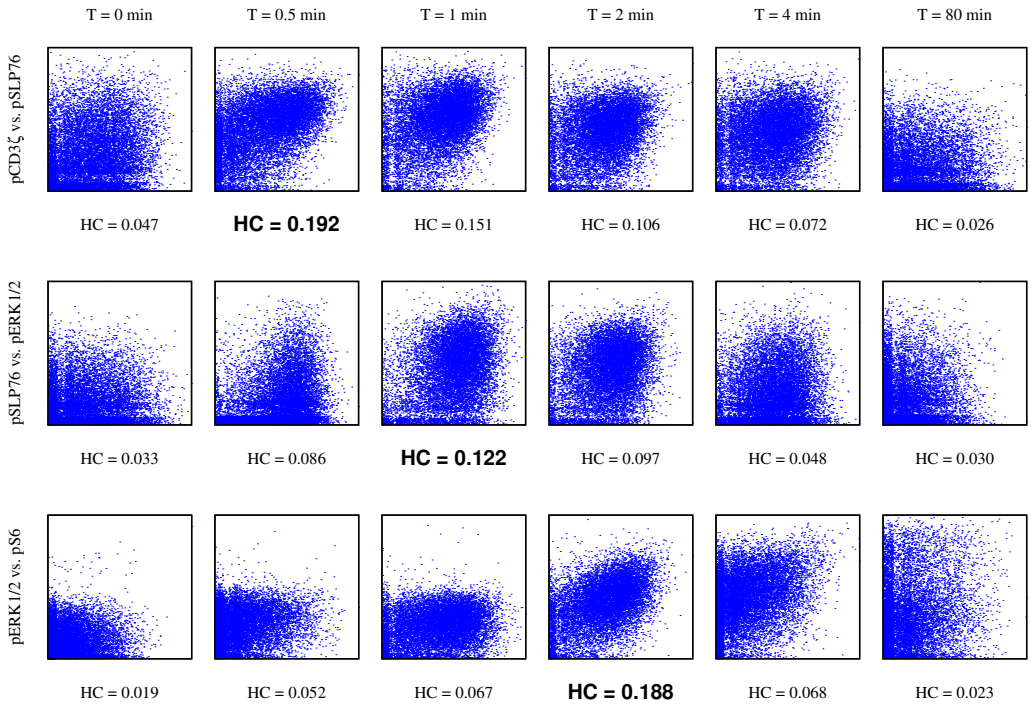

Figure 18. Scatter plots of gene pathway data for various pair of data and various time points (regular T-cells).

In Figure 19, the similar plots was shown for T-cells exposed with an antigen. Similarly, hypercontractivity is able to capture the trend.

We subsample the raw data from [7] to evaluate the ability to find the trend from smaller samples. Precisely, given a resampling rate $\gamma \in (0, 1]$, we randomly select a subset of indices $S_i \subseteq [n_i]$ with $\mathrm{card}(S_i) = \lceil \gamma n_i \rceil$, compute $\tau(X_{t_i}, Y_{t_i})$, $\tau(Y_{t_i}, Z_{t_i})$ and $\tau(Z_{t_i}, W_{t_i})$ from subsamples $\{X_{t_i}^{(j)}, Y_{t_i}^{(j)}, Z_{t_i}^{(j)}, W_{t_i}^{(j)}\}_{j \in S_i}$, and determine whether we can recover the trend successfully, i.e., whether $\arg\max_{t_i} \tau(X_{t_i}, Y_{t_i}) \leq \arg\max_{t_i} \tau(Y_{t_i}, Z_{t_i}) \leq \arg\max_{t_i} \tau(Z_{t_i}, W_{t_i})$. We repeat the experiment several times with independent subsamples and compute the probability of successfully recovering the trend. Figure 20 illustrates that when the entire dataset is available, all methods are able to recover the trend correctly. When only fewer samples are available, hypercontractivity improves upon other competing measures in recovering the hidden chronological order of interactions of the pathway. For completeness, we run datasets for both regular T-cells (shown in left figure) and T-cells exposed

with an antigen (shown right figure), for which we expect distinct biological trends. Hypercontractivity method can capture the trend for both datasets correctly and sample-efficiently.

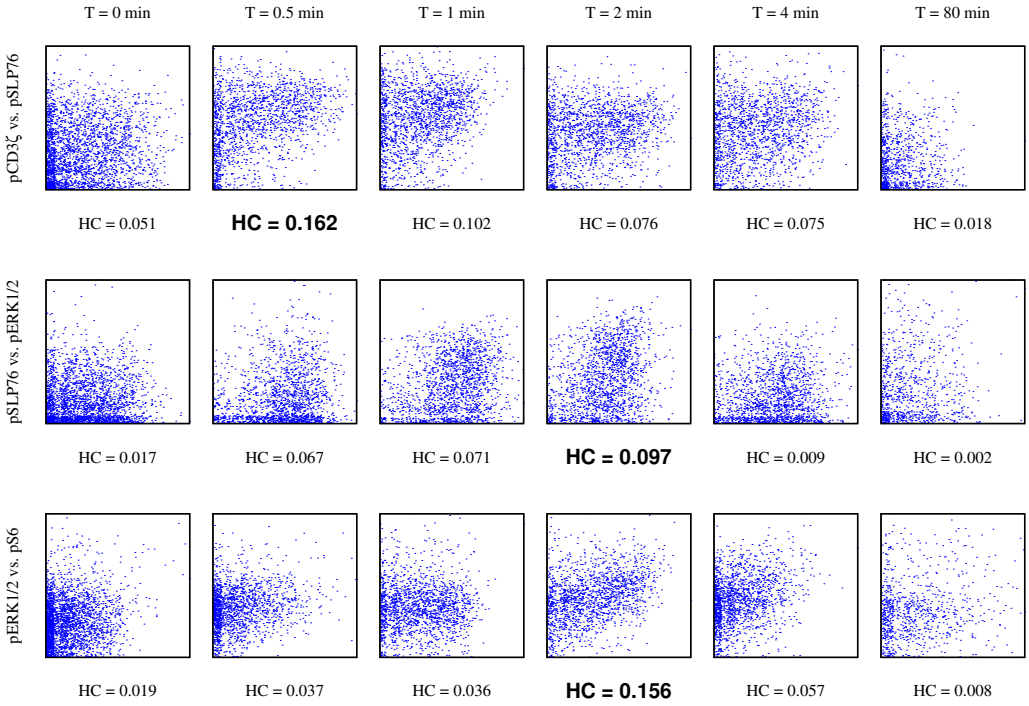

Figure 19. Scatter plots of gene pathway data for various pair of data and various time points (T-cells exposed with an antigen).

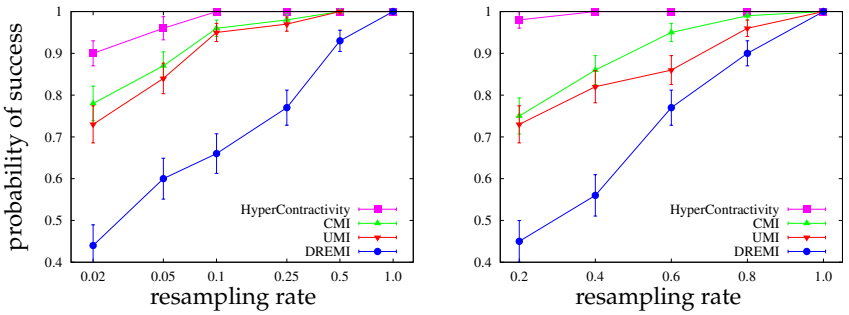

Figure 20. Accuracy vs. subsampling rate. Hypercontractivity method has higher probability to recover the trend when data size is smaller compared to other methods. (**left**) regular T-cells; (**right**) T-cells exposed with an antigen [7].

5. Proofs

In this section, we provide proofs for our main results and technical lemmas.

5.1. Proof of Proposition 1

Let $S_F(X,Y) = F(\sqrt{s(X;Y)}, \sqrt{s(Y;X)})$ for F satisfying conditions in Proposition 1. We show that $S_F(X,Y)$ satisfies all Rényi's axioms, i.e., Axioms 1–5 and 6′ and 7′.

1. $S_F(X,Y)$ is defined for any pair of non-constant random variables X, Y because $s(X;Y) \in [0,1]$ and $s(Y;X) \in [0,1]$ are defined for any random variables X, Y by Theorem 1.
2. $S_F(X,Y) \in [0,1]$ because the output of a function F is in $[0,1]$ by the condition on F.

3. If X and Y are statistically independent, $s(X;Y) = s(Y;X) = 0$. By the condition on F, it follows that $S_F(X,Y) = 0$. If $S_F(X,Y) = 0$, by the condition on F, $s(X;Y)s(Y;X) = 0$, which implies that X and Y are statistically independent.
4. $S_F(f(X),g(Y)) = S_F(X,Y)$ for any bijective Borel-measurable functions f, g because $\sqrt{s(f(X);g(Y))} = \sqrt{s(X;Y)}$ and $\sqrt{s(g(Y);f(X))} = \sqrt{s(Y;X)}$ by Theorem 1.
5. For $(X,Y) \sim \mathcal{N}(\mu,\Sigma)$ with Pearson correlation ρ, $s(X;Y) = s(Y;X) = \rho^2$. Hence, $S_F(X,Y) = F(|\rho|,|\rho|) = |\rho|$.

6′ If $Y = f(X)$ for a non-constant function f, it follows that $I(f(X);f(X)) = I(f(X);X)$ because if $f(X)$ is discrete, $I(f(X);f(X)) = I(f(X);X) = H(f(X))$ and otherwise, $I(f(X);f(X)) = I(f(X);X) = \infty$. Hence

$$s(X;f(X)) = \sup_{U-X-f(X)} I(U;f(X))/I(U;X) = I(f(X);f(X))/I(f(X);X) = 1.$$

Similarly, $s(f(X);X) = \sup_{U-f(X)-X} I(U;X)/I(U;f(X)) = 1$. Hence, $S_F(X;f(X)) = F(1,1) = 1$. Likewise, we can show that $S_F(X;Y) = 1$ if $X = g(Y)$.

7′ $S_F(X,Y) = S_F(Y,X)$ because $F(x,y) = F(y,x)$.

5.2. Proof of Theorem 1

We show that $s(X;Y)$ satisfies Axioms 1–6 in Section 2.

1. For any non-constant random variable X, $\exists\, U$ s.t. $I(U;X) > 0$. Hence, $s(X;Y)$ is defined for any pair of non-constant random variables X and Y.
2. Since mutual information is non-negative, $s(X;Y) \geq 0$. By data processing inequality, for any $U - X - Y$, $I(U;X) \leq I(U;Y)$. Hence, $s(X;Y) \leq 1$.
3. If X and Y are independent, for any U, $I(U;Y) \leq I(X;Y) = 0$. Hence, $s(X;Y) = 0$. If X and Y are dependent, $I(X;Y) > 0$, which implies that $s(X;Y) \geq I(X;Y)/H(X) > 0$.
4. For any bijective functions f, g,

$$I(U;g(Y)) = I(U;g(Y),Y) = I(U;Y) + I(U;g(Y)|Y) = I(U;Y).$$

Similarly, $I(U;f(X)) = I(U;X)$. Hence,

$$\begin{aligned} s(f(X);g(Y)) &= \sup_{U:U-f(X)-g(Y),I(U;f(X))>0} \frac{I(U;g(Y))}{I(U;f(X))} \\ &= \sup_{U:U-X-f(X)-g(Y)-Y,I(U;X)>0} \frac{I(U;Y)}{I(U;X)} \\ &= s(X;Y). \end{aligned}$$

5. By Theorem 3.1 in [33], for (X,Y) jointly Gaussian with correlation coefficient ρ,

$$\min_{U:\, U-X-Y} (I(U;X) - \beta I(U;Y)) = 0$$

for $\beta \leq 1/\rho^2$. Equivalently,

$$\max_{U:\, U-X-Y} \Big(I(U;Y) - \rho^2 I(U;X) \Big) = 0,$$

which implies that $s(X;Y) \leq \rho^2$. To show that $s(X;Y) \geq \rho^2$, let $U_Z = X + Z$ for $Z \sim (0,\sigma_1^2)$. Consider

$$\begin{aligned} s(X;Y) &\geq \lim_{\sigma_1^2 \to \infty} \frac{I(U_Z;Y)}{I(U_Z;X)} \\ &= \lim_{\sigma_1^2 \to \infty} \frac{\log\left(\frac{(\sigma_X^2+\sigma_1^2)\sigma_Y^2}{(\sigma_X^2+\sigma_1^2)\sigma_Y^2-\rho^2\sigma_X^2\sigma_Y^2}\right)}{\log\left(1+\frac{\sigma_X^2}{\sigma_1^2}\right)} \\ &= \lim_{\sigma_1^2 \to \infty} \frac{\rho^2\sigma_X^2\sigma_Y^2 / \left((\sigma_X^2+\sigma_1^2)\sigma_Y^2-\rho^2\sigma_X^2\sigma_Y^2\right)}{\sigma_X^2/\sigma_1^2} \\ &= \rho^2. \end{aligned}$$

Hence, $s(X;Y) = \rho^2$. An alternative proof is provided in [24].

6. To prove that $s(X;Y)$ satisfies Axiom 6, we first show the following lemma.

Lemma 1. *Consider a pair of random variables $(X,Y) \in \mathcal{X} \times \mathcal{Y}$. The hypercontractivity $s(X;Y)$ is lower bounded by*

$$s(X;Y) \geq \frac{I(U;Y|X \in \mathcal{X}_r)}{H(\alpha)/\alpha + I(U;X|X \in \mathcal{X}_r)} \tag{12}$$

for any $\mathcal{X}_r$ such that $\mathcal{X}_r \subseteq \mathcal{X}$ for $P\{X \in \mathcal{X}_r\} =: \alpha > 0$.

Proof. Let

$$U_s = \begin{cases} U \sim p(u|x) & \text{if } X \in \mathcal{X}_r, \\ \emptyset & \text{otherwise.} \end{cases} \tag{13}$$

Let $S = \mathbb{I}_{\{U_s=\emptyset\}} = \mathbb{I}_{\{X \in \mathcal{X}_r\}}$. Note that $S - U_s - X - Y$ holds, and that S is a deterministic function of X. Hence,

$$\begin{aligned} I(U_s;X) &= I(U_s,S;X) \\ &= I(S;X) + I(U_s;X|S) \\ &= H(\alpha) + \alpha I(U;X|X \in \mathcal{X}_r). \end{aligned} \tag{14}$$

Consider

$$\begin{aligned} I(U_s;Y) &= I(U_s,S;Y) \\ &= I(S;Y) + I(U_s;Y|S) \\ &\geq \alpha I(U;Y|X \in \mathcal{X}_r). \end{aligned} \tag{15}$$

The proof is completed by combining (14) and (15). □

Assume that $Y = f(X)$ for $X \in \mathcal{X}_r$. Considering $U = f(X)$ in (13) in Lemma 1, we obtain the following lower bound:

$$s(X;Y) \geq \frac{I(f(X);f(X)|X \in \mathcal{X}_r)}{H(\alpha)/\alpha + I(f(X);X|X \in \mathcal{X}_r)}.$$

For any continuous random variable X and a non-constant continuous function f, $I(f(X);f(X)|X \in \mathcal{X}_r) = I(f(X);X|X \in \mathcal{X}_r) = \infty$, which implies that $s(X;Y) = 1$.

5.3. Proof of Theorem 2

We first prove that $\mathrm{mCor}(X,Y) = \sqrt{\alpha}\,\mathrm{mCor}(X_r,Y)$ in (2). Let $S = \mathbb{I}_{\{X\in\mathcal{X}_r\}}$ be the indicator for whether $X \in \mathcal{X}_r$ or not. Consider

$$\begin{aligned}
\mathrm{mCor}(X;Y) &= \max_{\substack{f,g \\ :\mathsf{E}[f(X)]=\mathsf{E}[g(Y)]=0, \\ \mathsf{E}[f^2(X)]\le 1,\mathsf{E}[g^2(Y)]\le 1}} \mathsf{E}[f(X)g(Y)] \\
&= \max_{\substack{f,g \\ :\mathsf{E}[f(X)]=\mathsf{E}[g(Y)]=0, \\ \mathsf{E}[f^2(X)]\le 1,\mathsf{E}[g^2(Y)]\le 1}} \mathsf{E}_S[\mathsf{E}[f(X)g(Y)\,|\,S]] \\
&= \max_{\substack{f,g \\ :\mathsf{E}[f(X)]=\mathsf{E}[g(Y)]=0, \\ \mathsf{E}[f^2(X)]\le 1,\mathsf{E}[g^2(Y)]\le 1}} (\alpha\,\mathsf{E}[f(X)g(Y)\,|\,X\in\mathcal{X}_r] + \bar{\alpha}\,\mathsf{E}[f(X)g(Y)\,|\,X\in\mathcal{X}_d]) \\
&= \max_{\substack{f,g \\ :\mathsf{E}[f(X)]=\mathsf{E}[g(Y)]=0, \\ \mathsf{E}[f^2(X)]\le 1,\mathsf{E}[g^2(Y)]\le 1}} (\alpha\,\mathsf{E}[f(X)g(Y)\,|\,X\in\mathcal{X}_r] + \bar{\alpha}\,\mathsf{E}[f(X)\,|\,X\in\mathcal{X}_d]\,\mathsf{E}[g(Y)\,|\,X\in\mathcal{X}_d]) \\
&\overset{(a)}{=} \alpha \max_{\substack{f,g \\ :\mathsf{E}[f(X)]=\mathsf{E}[g(Y)]=0, \\ \mathsf{E}[f^2(X)]\le 1,\mathsf{E}[g^2(Y)]\le 1}} \mathsf{E}[f(X)g(Y)\,|\,X\in\mathcal{X}_r] \\
&\overset{(b)}{=} \sqrt{\alpha}\,\mathrm{mCor}(X_r,Y).
\end{aligned}$$

Step (a) holds since $\mathsf{E}[g(Y)|X\in\mathcal{X}_r] = \mathsf{E}[g(Y)|X\in\mathcal{X}_d]$ from the assumption that marginal distributions are equal, and that $\mathsf{E}[g(Y)] = \alpha\,\mathsf{E}[g(Y)|X\in\mathcal{X}_r] + \bar{\alpha}\,\mathsf{E}[g(Y)|X\in\mathcal{X}_d]$. To show step (b), let $c = \mathsf{E}[f(X)|X\in\mathcal{X}_d]$ and note that

$$\begin{aligned}
\alpha\,\mathsf{E}[f(X)\,|\,X\in\mathcal{X}_r] &= -\bar{\alpha}c, \\
\alpha\,\mathsf{E}[f^2(X)\,|\,X\in\mathcal{X}_r] &= \mathsf{E}[f^2(X)] - \bar{\alpha}\,\mathsf{E}[f^2(X)\,|\,X\in\mathcal{X}_d] \\
&\le 1-\bar{\alpha}c^2, \\
\mathsf{E}[g(Y)\,|\,X\in\mathcal{X}_r] &= 0.
\end{aligned}$$

Hence,

$$\begin{aligned}
\max_{\substack{f,g \\ :\mathsf{E}[f(X)]=\mathsf{E}[g(Y)]=0, \\ \mathsf{E}[f^2(X)]\le 1,\mathsf{E}[g^2(Y)]\le 1}} \mathsf{E}[f(X)g(Y)\,|\,X\in\mathcal{X}_r] &= \max_{\substack{f_r,g \\ :\mathsf{E}[f_r(X)]=-\bar{\alpha}c/\alpha,\mathsf{E}[g(Y)]=0, \\ \mathsf{E}[f_r^2(X)]\le(1-\bar{\alpha}c^2)/\alpha,\mathsf{E}[g^2(Y)]\le 1}} \mathsf{E}[f_r(X)g(Y)] \\
&= \max_{\substack{f_{rc},g \\ :\mathsf{E}[f_{rc}(X)]=0,\mathsf{E}[g(Y)]=0, \\ \mathsf{E}[f_{rc}^2(X)]\le(\alpha-\bar{\alpha}c^2)/\alpha^2,\mathsf{E}[g^2(Y)]\le 1}} \mathsf{E}[(f_{rc}(X)g(Y)] \\
&= \max_{\substack{f_{rc},g \\ :\mathsf{E}[f_{rc}(X)]=0,\mathsf{E}[g(Y)]=0, \\ \mathsf{E}[f_{rc}^2(X)]\le 1/\alpha,\mathsf{E}[g^2(Y)]\le 1}} \mathsf{E}[f_{rc}(X)g(Y)] \\
&= \max_{\substack{f_{rca},g \\ :\mathsf{E}[f_{rca}(X)]=0,\mathsf{E}[g(Y)]=0, \\ \mathsf{E}[f_{rca}^2(X)]\le 1,\mathsf{E}[g^2(Y)]\le 1}} \frac{1}{\sqrt{\alpha}}\,\mathsf{E}[f_{rca}(X)g(Y)] \\
&= \frac{\mathrm{mCor}(X_r,Y)}{\sqrt{\alpha}},
\end{aligned}$$

where $f_r(X)$, $f_{rc}(X) = f_r(X) + \bar{\alpha}c/\alpha$, and $f_{rca}(X) = \sqrt{\alpha}f_{rc}(X)$ are functions defined only for $X\in\mathcal{X}_r$.

We next show $\mathrm{dCor}(X,Y) = \alpha\ \mathrm{dCor}(X_r,Y)$ in (3). Let

$$h_X(s) = \mathsf{E}[e^{isX}],\ h_Y(t) = \mathsf{E}[e^{itY}],\ h_{XY}(s,t) = \mathsf{E}[e^{i(sX+tY)}].$$

Note that

$$\begin{aligned} h_{XY}(s,t) &= \mathsf{E}[e^{i(sX+tY)}] \\ &= \alpha\,\mathsf{E}[e^{i(sX+tY)}|X\in\mathcal{X}_r] + \bar{\alpha}\,\mathsf{E}[e^{isX}|X\in\mathcal{X}_d]\,\mathsf{E}[e^{itY}|X\in\mathcal{X}_d] \\ &= \alpha\,\mathsf{E}[e^{i(sX+tY)}|X\in\mathcal{X}_r] + \bar{\alpha}\,\mathsf{E}[e^{isX}|X\in\mathcal{X}_d]\,\mathsf{E}[e^{itY}], \end{aligned} \tag{16}$$

and

$$h_X(s) = \mathsf{E}[e^{isX}] = \alpha\,\mathsf{E}[e^{isX}|X\in\mathcal{X}_r] + \bar{\alpha}\,\mathsf{E}[e^{isX}|X\in\mathcal{X}_d]. \tag{17}$$

By combining (16) and (17),

$$\begin{aligned} h_{XY}(s,t) - h_X(s)h_Y(t) &= \alpha\,\mathsf{E}[e^{i(sX+tY)}|X\in\mathcal{X}_r] - \alpha\,\mathsf{E}[e^{isX}|X\in\mathcal{X}_r]\,\mathsf{E}[e^{itY}] \\ &= \alpha\,\mathsf{E}[e^{i(sX+tY)}|X\in\mathcal{X}_r] - \alpha\,\mathsf{E}[e^{isX}|X\in\mathcal{X}_r]\,\mathsf{E}[e^{itY}|X\in\mathcal{X}_r] \\ &= \alpha\ \mathrm{dCor}(X_r,Y). \end{aligned}$$

Finally, we show that $\mathrm{MIC}(X,Y) \le \alpha\ \mathrm{MIC}(X_r,Y)$ in (4).

Let $X_Q(X) \in \mathcal{X}_Q(X)$ and $Y_Q(Y) \in \mathcal{Y}_Q(Y)$ denote a quantization of X and Y, respectively. Consider

$$\begin{aligned} \mathrm{MIC}(X,Y) &= \max_{X_Q(X),Y_Q(Y)} \frac{I(X_Q;Y_Q)}{\log\min\{|\mathcal{X}_Q|,|\mathcal{Y}_Q|\}} \\ &\le \max_{X_Q(X),Y_Q(Y)} \frac{I(\mathbb{I}_{X\in\mathcal{X}_r},X_Q;Y_Q)}{\log\min\{|\mathcal{X}_Q|,|\mathcal{Y}_Q|\}} \\ &\overset{(a)}{=} \alpha \max_{X_Q(X),Y_Q(Y)} \frac{I(X_Q;Y_Q|X\in\mathcal{X}_r)}{\log\min\{|\mathcal{X}_Q|,|\mathcal{Y}_Q|\}} \\ &\le \alpha \max_{X_Q(X_r),Y_Q(Y)} \frac{I(X_Q;Y_Q|X\in\mathcal{X}_r)}{\log\min\{|\mathcal{X}_Q(X_r)|,|\mathcal{Y}_Q|\}} \\ &= \alpha\ \mathrm{MIC}(X_r,Y), \end{aligned}$$

where step (a) holds because $\mathbb{I}_{X\in\mathcal{X}_r} \perp\!\!\!\perp Y$ implies $\mathbb{I}_{X\in\mathcal{X}_r} \perp\!\!\!\perp Y_Q$ and $X \perp\!\!\!\perp Y$ in $X\in\mathcal{X}_d$ implies $X_Q \perp\!\!\!\perp Y_Q$ in $X\in\mathcal{X}_d$.

5.4. Proof of Proposition 2

The inverse hypercontractivity $s(Y;X)$ is defined as

$$s(Y;X) = \sup_{U-Y-X} \frac{I(U;X)}{I(U;Y)}.$$

Let $\mathbb{I}_r = \mathbb{I}_{\{X\in\mathcal{X}_r\}}$. Since the marginal distribution of Y given $\{X\in\mathcal{X}_r\}$ and the one given $\{X\notin\mathcal{X}_r\}$ are equivalent, Y and $\mathbb{I}_r$ are independent, i.e., $I(Y;\mathbb{I}_r) = 0$. For any U such that Markov chain $U-Y-X$ holds, the Markov chain $U-Y-X-\mathbb{I}_r$ holds. Hence, $I(U;\mathbb{I}_r) = 0$. Hence, for any $U-Y-X$, consider

$$\begin{aligned} I(U;X) &= I(U;X,\mathbb{I}_r) \\ &= I(U;X|\mathbb{I}_r) \\ &= (1-\alpha)I(U;X|\mathbb{I}_r=0) + \alpha I(U;X|\mathbb{I}_r=1) \\ &\overset{(a)}{=} \alpha I(U;X|\mathbb{I}_r=1) \end{aligned}$$

Step (a) holds because $Y \perp\!\!\!\perp X$ given $\mathbb{I}_r = 0$. Consider

$$\begin{aligned} I(U;Y) &\overset{(a)}{=} I(U;Y,\mathbb{I}_r) \\ &= I(U;Y|\mathbb{I}_r) + I(U;\mathbb{I}_r) \\ &\overset{(b)}{=} I(U;Y|\mathbb{I}_r) \\ &= \alpha I(U;Y|\mathbb{I}_r=1) + (1-\alpha)I(U;Y|\mathbb{I}_r=0) \\ &\overset{(c)}{=} I(U;Y|\mathbb{I}_r=1), \end{aligned}$$

where step (a) folllows since $U - Y - \mathbb{I}_r$. Step (b) follows from $I(U;\mathbb{I}_r) = 0$. Step (c) holds since $H(U|\mathbb{I}_r = 1) = H(U|\mathbb{I}_r = 0)$ and $U - Y - \mathbb{I}_r$. Therefore, for any $U - Y - X$, it follows that

$$s(Y;X) = \sup_{U-Y-X} \frac{\alpha I(U;X|I_r=1)}{I(U;Y|I_r=1)} = \alpha\, s(Y;X_r).$$

5.5. Noisy Rare Correlation in Example 1

Let

$$U = X + Z, \;\; Z \sim \mathcal{N}(0,\sigma_1^2).$$

Consider

$$\sup_{U:U-X-Y, I(U;X)>0} \frac{I(U;Y)}{I(U;X)} \geq \sup_{\sigma_1^2 \geq 0} \frac{\log \frac{(1+\sigma_1^2)}{(1+\sigma_1^2)-\rho^2}}{H(\alpha)/\alpha + \log(1+1/\sigma_1^2)}.$$

The inequality (7) follows by choosing $\sigma_1^2 = (1-\rho^2)/\rho^2$.

5.6. Noisy Discrete Rare Correlation in Example 3

The inequality (8) follows by choosing $r(x) = \mathbb{I}_{\{X=1\}}$ in (10). To show (9), we show that

$$\mathrm{mCor}(X_r, Y) = 1 - \frac{k}{k-1}\epsilon. \tag{18}$$

The rest follows because $\mathrm{mCor}(X;Y) = \sqrt{\alpha}\,\mathrm{mCor}(X_r,Y)$ by Proposition 2. To show (18), we use the fact that maximal correlation is the second eigenvalue of $Q = P_X^{-1/2} P_{XY} P_Y^{-1/2}$ (see [40] for a detailed proof). We can easily show that

$$Q = \left(1 - \frac{k}{k-1}\epsilon\right) I + \frac{\epsilon}{k-1} 1 1^T.$$

First singular vector of Q is $P_X^{1/2} = 1/\sqrt{k}$. Second singular vector u_2 is orthogonal to $1/\sqrt{k}$. The Equation (18) follows because $\mathrm{mCor}(X_r;Y) = u_2^T Q u_2 = u_2^T(1 - k\epsilon/(k-1))u_2$.

5.7. Proof of Proposition 4

We first prove the second part of proposition: the hypercontractivity coefficient $\sqrt{s(X;Y)}$ satisfies Axioms 1–4, 5′, and 6. It follows immediately from Theorem 1 that $\sqrt{s(X;Y)}$ satisfies Axioms 1–4 and 6 because in the proof of Theorem 1—1–4 and 6, the same argument holds for random vectors

X and Y. We can show that that $\sqrt{s(X;Y)}$ satisfies Axiom 5′ using results from [33]. In [33], it is shown that that as we increase β starting from zero, $\min\{I(U;X) - \beta I(U;Y)\}$ departs form zero at $\beta = 1/\|\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\|^2$ for jointly Gaussian random vectors X and Y. This result implies that $\sqrt{s(X;Y)} = \|\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\|$.

To show that maximal correlation of two random vectors satisfies Axioms 1–4, 6′, and 7′, we can follow the same arguments for showing that maximal correlation for two random variables satisfies Axioms 1–4, 6′, and 7′ by [4]. To show that maximal correlation satisfies Axiom 5′, note that maximal correlation is upper bounded by hypercontractivity as shown in Remark 1 in Section 2.3: hence $\text{mCor}(X;Y) \leq \|\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\|$ for a jointly Gaussian X, Y. Equality holds because $\text{mCor}(X,Y)$ is lower bounded by its canonical correlation, which is $\|\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\|$ for jointly Gaussian random vectors (X,Y) [33].

5.8. Proof of Theorem 3

We begin with the following assumptions:

(a) There exist finite constants $C_1 < C_1' < C_2' < C_2$ such that the ratio of the optimal r_x^* and the true p_x satisfies $r_x^*(x)/p_x(x) \in [C_1', C_2']$ for every $x \in \mathcal{X}$.
(b) There exist finite constants $C_0' > C_0 > 0$ such that the KL divergence $D(r_x^*||p_x) > C_0'$.

With a little abuse of notations, we define $s(r_x) = D(r_y||p_y)/D(r_x||p_x)$ and $\widehat{s}(\mathbf{w}) = \mathbf{w}^T\mathbf{A}\log(\mathbf{A}^T\mathbf{w})/\mathbf{w}^T\log\mathbf{w}$. Therefore, $s(X;Y) = \max_{r_x \in R} s(r_x)$ and $\widehat{s}_\Delta(X;Y) = \max_{\mathbf{w}\in T_\Delta}\widehat{s}(\mathbf{w})$. Here R is the probability simplex over all r_x. We want to bound the error $|\widehat{s}_\Delta(X;Y) - s(X;Y)|$. First, consider the quantity:

$$s_\Delta(X;Y) \equiv \max_{r_x \in T_\Delta(R)} s(r_x)\,, \tag{19}$$

where the constraint set $T_\Delta(R)$ is defined as:

$$T_\Delta(R) = \{r_x \in \mathbb{R}^{|\mathcal{X}|} : [(r_x(x)/p_x(x))] \in T_\Delta \text{ and } \sum_{x\in\mathcal{X}} r_x(x) \in [1 - |\mathcal{X}|\Delta, 1 + |\mathcal{X}|\Delta]\} \tag{20}$$

Now we rewrite the error term as

$$\left|\widehat{s}_\Delta(X;Y) - s(X;Y)\right| \leq |s_\Delta(X;Y) - s(X;Y)| + |\widehat{s}_\Delta(X;Y) - s_\Delta(X;Y)|\,. \tag{21}$$

The first error comes from quantization. Let r^* be the maximizer of $s(X;Y)$. By assumption, $r^*(x)/p_x(x) \in [C_1, C_2]$, for all x. Since $T_\Delta(R)$ is a quantization of the simplex R, so there exists an $r_0 \in T_\Delta(R)$ such that $|r_0(x) - r^*(x)| < \Delta$ for all $x \in \mathcal{X}$. Now we will bound the difference between $s(r_0)$ and $s(r^*)$ by the following lemma:

Lemma 2. *If $r(x)/p(x) \in [C_1, C_2]$ and $r'(x)/p(x) \in [C_1, C_2]$ for all $x \in \mathcal{X}$, and $D(r_x||p_x) > C_0$ and $D(r_x'||p_x) > C_0$, then*

$$\left|s(r) - s(r')\right| \leq L\max_{x\in\mathcal{X}}|r(x) - r'(x)|\,, \tag{22}$$

for some positive constant L.

Next we have:

$$\begin{aligned} s(X;Y) &= s(r^*) \leq s(r_0) + L\max_{x\in\mathcal{X}}|r_0(x) - r^*(x)| \\ &\leq \max_{r\in T_\Delta(R)} s(r) + L\Delta = s_\Delta(X;Y) + L\Delta. \end{aligned} \tag{23}$$

Similarly, let r^{**} be the maximizer of $s_\Delta(X;Y)$, we can also find a $r_1 \in R$ such that $|r_1(x) - r^{**}(x)| < \Delta$ for all $x \in \mathcal{X}$. Using Lemma 2 again, we will obtain $s_\Delta(X;Y) \leq s(X;Y) + L\Delta$. Therefore, the quantization error is bounded by $O(\Delta)$ with probability 1.

Now consider the second term. Upper bound on the second term relies on the convergence of estimation of s. We claim that for given r_x, the estimator is convergent in probability, i.e.,

Lemma 3.

$$\lim_{N\to\infty} \mathcal{P}\left(\left| \widehat{s}(\mathbf{w}_r) - s(r_x) \right| > \varepsilon \right) = 0 \,. \tag{24}$$

Here $\mathbf{w}_r(x) = r_x(x)/p_x(x)$. Since the set $T_\Delta(R)$ is finite, by union bound, we have:

$$\begin{aligned} &\lim_{N\to\infty} \mathcal{P}\big(\forall r \in T_\Delta(R), \left| \widehat{s}(\mathbf{w}_r) - s(r_x) \right| \leq \varepsilon \big) \\ &\geq 1 - |T_\Delta(R)| \lim_{N\to\infty} \mathcal{P}\left(\left| \widehat{s}(\mathbf{w}_r) - s(r_x) \right| \leq \varepsilon \right) = 1. \end{aligned} \tag{25}$$

Also, by the strong law of large numbers, we have that

$$\lim_{N\to\infty} \mathcal{P}\left(\forall x \in \mathcal{X}, |p_x(x) - \frac{n_x}{n}| < \frac{\Delta}{C_2|\mathcal{X}|} \right) = 1. \tag{26}$$

where $n_x = \text{card}\{i \in [n] : x_i = x\}$. We claim that if the events inside the probability in (25) and (26) happen simultaneously, then $|\widehat{s}_\Delta(X;Y) - s_\Delta(X;Y)| < \varepsilon + O(\Delta)$, which implies the desired claim.

Let $\mathbf{w}^* = \arg\max_{\mathbf{w}\in T_\Delta} \widehat{s}(\mathbf{w})$. Define $r_2(x) = \mathbf{w}^*(x)p_x(x)$. Since $[r_2(x)/p_x(x)] \in T_\Delta$ for all x and

$$\begin{aligned} \left| \textstyle\sum_{x\in\mathcal{X}} r_2(x) - 1 \right| &= \left| \textstyle\sum_{x\in\mathcal{X}} \mathbf{w}^*_x (p_x(x) - \frac{n_x}{n}) \right| + \frac{\Delta|\mathcal{X}|}{2} \\ &\leq |\mathcal{X}| \left(\tfrac{\Delta}{2} + C_2 \max_{x\in\mathcal{X}} \left| p_x(x) - \tfrac{n_x}{n} \right| \right) \\ &\leq (|\mathcal{X}|/2 + 1)\Delta \,. \end{aligned} \tag{27}$$

Therefore, $r_2 \in T_\Delta(R)$, so

$$\widehat{s}_\Delta(X;Y) = \widehat{s}(\mathbf{w}^*) \leq s(r_2) + \varepsilon \leq s_\Delta(X;Y) + \varepsilon. \tag{28}$$

On the other hand, consider $r^{**} = \arg\max_{r_x \in T_\Delta(R)} s(r_x)$ again, and define $\mathbf{w}_0(x) = r^{**}(x)/p_x(x)$. We know that $\mathbf{w}_0 \in T_\Delta^{|\mathcal{X}|}$ but not necessarily $\sum_{i=1}^n \mathbf{w}_0(x_i) = n$. However, we claim that the sum is closed to n as follows:

$$\begin{aligned} \left| \textstyle\sum_{i=1}^n \mathbf{w}_0(x_i) - n \right| &= \left| \textstyle\sum_{x\in\mathcal{X}} \frac{n_x r^{**}(x)}{p_x(x)} - n \right| \\ &\leq n \max_{x\in\mathcal{X}} \left\{ \tfrac{r^{**}(x)}{p_x(x)} \left| \tfrac{n_x}{n} - p_x(x) \right| \right\} \\ &\leq nC_2 \tfrac{\Delta}{C_2|\mathcal{X}|} < n\Delta \end{aligned} \tag{29}$$

so we can find a $\mathbf{w}_1 \in T_\Delta(R)$ such that $|\mathbf{w}_1(x) - \mathbf{w}_0(x)| \leq \Delta$ for all x. Let $r_4(x) = \mathbf{w}_1(x)p_x(x)$, similar as (27), we know that $r_4 \in T_\Delta(R)$. Moreover, $\left| r_4(x) - r^{**}(x) \right| \leq p_x(x) \left| \mathbf{w}_1(x) - \mathbf{w}_0(x) \right| \leq \Delta$ for all x. Then we have

$$\begin{aligned} s_\Delta(X;Y) &= s(r^{**}) \leq s(r_4) + L \max_{x\in\mathcal{X}} |r^{**}(x) - r_4(x)| \\ &\leq \widehat{s}(\mathbf{w}_1) + \varepsilon + L\Delta = \widehat{s}_\Delta(X;Y) + \varepsilon + L\Delta. \end{aligned} \tag{30}$$

We conclude that $|\widehat{s}_\Delta(X;Y) - s_\Delta(X;Y)| < \varepsilon + O(\Delta)$; thus our proof is complete.

5.9. Proof of Lemma 2

We will show that for any $x \in \mathcal{X}$, we have $|\partial s(r_x)/\partial r_x(x)| \leq L/|\mathcal{X}|$ for some L. Therefore,

$$\begin{array}{rl} \left| s(r) - s(r') \right| & \leq \sum_{x\in\mathcal{X}} |\frac{\partial s(r)}{\partial r_x(x)}|\, |r_x(x) - r'_x(x)| \\ & \leq L \max_{x\in\mathcal{X}} |r_x(x) - r'_x(x)| \,. \end{array} \tag{31}$$

The gradient can be written as

$$\begin{array}{rl} \frac{\partial s(r)}{\partial r_x(x)} & = \frac{\partial}{\partial r_x(x)} \frac{D(r_y||p_y)}{D(r_x||p_x)} \\ & = \frac{1}{D^2(r_x||p_x)} \left(\frac{\partial D(r_y||p_y)}{\partial r_x(x)} D(r_x||p_x) - \frac{\partial D(r_x||p_x)}{\partial r_x(x)} D(r_y||p_y) \right) . \end{array} \tag{32}$$

Since

$$\begin{array}{rcl} \frac{\partial D(r_x||p_x)}{\partial r_x(x)} & = & \log \frac{r_x(x)}{p_x(x)} + 1 \leq \max\{|\log C_1|, |\log C_2|\} + 1 \\ \frac{\partial D(r_y||p_y)}{\partial r_x(x)} & = & \int \frac{\partial r_y(y)}{\partial r_x(x)} \frac{\partial D(r_y||p_y)}{\partial r_y(y)} dy \\ & = & \int p_{y|x}(y|x)(\log \frac{r_y(y)}{p_y(y)} + 1) dy \leq \max\{|\log C_1|, |\log C_2|\} + 1 \end{array} \tag{33}$$

Therefore, we have

$$\begin{array}{rcl} |\frac{\partial s(r)}{\partial r_x(x)}| & \leq & (\max\{|\log C_1|, |\log C_2|\} + 1) \frac{D(p_x||r_x) + D(r_y||p_y)}{D^2(r_x||p_x)} \\ & \leq & \frac{2(\max\{|\log C_1|, |\log C_2|\}+1)}{D(r_x||p_x)} \\ & \leq & \frac{2(\max\{|\log C_1|, |\log C_2|\}+1)}{C_0} \end{array} \tag{34}$$

Since C_0, C_1, C_2 are constants and $|\mathcal{X}|$ is finite, our proof is complete by letting $L = 2|\mathcal{X}|(\max\{|\log C_1|, |\log C_2|\} + 1)/C_0$.

5.10. Proof of Lemma 3

Note that $\widehat{s}(\mathbf{w}_r) = \mathbf{w}^T \mathbf{A} \log(\mathbf{A}^T \mathbf{w}) / \mathbf{w}^T \log \mathbf{w}$. Define $\widehat{D}(r_y||p_y) = \mathbf{w}^T \mathbf{A} \log(\mathbf{A}^T \mathbf{w})$ and $\widehat{D}(r_x||p_x) = \mathbf{w}^T \log \mathbf{w}$. We will prove that both $\widehat{D}(r_y||p_y)$ converges to $D(r_y||p_y)$ and $\widehat{D}(r_x||p_x)$ converges to $D(r_x||p_x)$ in probability. Since $D(r_x||p_x) > 0$ and $\widehat{D}(r_x||p_x) > 0$ with probability 1, we obtain that $\widehat{s}(\mathbf{w}_r)$ converges to $D(r_y||p_y)/D(r_x||p_x) = s(r_x)$ in probability.

The convergence $\widehat{D}(r_x||p_x)$ comes from law of large number. Since $\widehat{D}(r_x||p_x) = \frac{1}{n} \sum_{i=1}^n \frac{r_x(X_i)}{p_x(X_i)} \log \frac{r_x(X_i)}{p_x(X_i)}$ and $D(r_x||p_x) = \mathbb{E}_{X\sim p_x} \left[\frac{r_x(X)}{p_x(X)} \log \frac{r_x(X)}{p_x(X)} \right]$, the weak law of large number shows the convergence in probability.

For the convergence of $\widehat{D}(r_y||p_y)$. Consider the vector $\mathbf{v} = \mathbf{A}^T \mathbf{w}$, we have

$$v_j = \frac{1}{n} \sum_{i=1}^n \frac{p_{xy}(X_i, Y_j)}{p_x(X_i) p_y(Y_j)} w_i = \frac{1}{n} \sum_{i=1}^n \frac{p_{y|x}(Y_j|X_i)}{p_y(Y_j)} \frac{r_x(X_i)}{p_x(X_i)}.$$

On the other hand, for fixed $Y_j = y$, we have

$$\frac{r_y(y)}{p_y(y)} = \frac{\mathbb{E}_{X\sim p_x} \left[p_{y|x}(y|X) \frac{r_x(X)}{p_x(X)} \right]}{p_y(y)} = \mathbb{E}_{X\sim p_x} \left[\frac{p_{y|x}(y|X)}{p_y(y)} \frac{r_x(X)}{p_x(X)} \right] .$$

Therefore, by law of large number, we conclude that v_j converges to $\frac{r_y(Y_j)}{p_y(Y_j)}$ in probability. Hence, $\widehat{D}(r_y||p_y) = \frac{1}{n} \sum_{j=1}^n v_j \log v_j$ converges to $\frac{1}{n} \sum_{j=1}^n \frac{r_y(Y_j)}{p_y(Y_j)} \log \frac{r_y(Y_j)}{p_y(Y_j)}$ in probability. Furthermore,

$\frac{1}{n}\sum_{j=1}^{n}\frac{r_y(Y_j)}{p_y(Y_j)}\log\frac{r_y(Y_j)}{p_y(Y_j)}$ converges to $D(r_y||p_y)=\mathbb{E}_{Y\sim p_y}\left[\frac{r_y(Y)}{p_y(Y)}\log\frac{r_y(Y)}{p_y(Y)}\right]$ in probability, by law of large number again. Therefore, we conclude that $\widehat{D}(r_y||p_y)$ converges to $D(r_y||p_y)$ in probability.

Acknowledgments: This work was partially supported by NSF grants CNS-1527754, CNS-1718270, CCF-1553452, CCF-1617745, CCF-1651236, CCF-1705007, and GOOGLE Faculty Research Award.

Author Contributions: All authors contributed equally to this work. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pearson, K. Note on Regression and Inheritance in the Case of Two Parents. *Proc. R. Soc. Lond.* **1895**, *58*, 240–242.
2. Hirschfeld, H. A Connection between Correlation and Contingency. *Math. Proc. Camb. Philos. Soc.* **1935**, *31*, 520–524.
3. Gebelein, H. Das statistische Problem der Korrelation als Variations-und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *J. Appl. Math. Mech.* **1941**, *21*, 364–379.
4. Rényi, A. On measures of dependence. *Acta Math. Hung.* **1959**, *10*, 441–451.
5. Reshef, D.N.; Reshef, Y.A.; Finucane, H.K.; Grossman, S.R.; McVean, G.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting Novel Associations in Large Data Sets. *Science* **2011**, *334*, 1518–1524.
6. Székely, G.J.; Rizzo, M.L.; Bakirov, N.K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **2007**, *35*, 2769–2794.
7. Krishnaswamy, S.; Spitzer, M.H.; Mingueneau, M.; Bendall, S.C.; Litvin, O.; Stone, E.; Pe'er, D.; Nolan, G.P. Conditional density-based analysis of T cell signaling in single-cell data. *Science* **2014**, *346*, 1250689.
8. Tishby, N.; Pereira, F.C.; Bialek, W. The Information Bottleneck Method. In Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing, Champaign, IL, USA, 22–24 September 1999; pp. 368–377.
9. Dhillon, I.S.; Mallela, S.; Kumar, R. A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification. *J. Mach. Learn. Res.* **2003**, *3*, 1265–1287.
10. Bekkerman, R.; El-Yaniv, R.; Tishby, N.; Winter, Y. Distributional Word Clusters vs. Words for Text Categorization. *J. Mach. Learn. Res.* **2003**, *3*, 1183–1208.
11. Anantharam, V.; Gohari, A.A.; Kamath, S.; Nair, C. On Maximal Correlation, Hypercontractivity, and the Data Processing Inequality studied by Erkip and Cover. *arXiv* **2013**, arXiv:1304.6133.
12. Davies, E.B.; Gross, L.; Simone, B. Hypercontractivity: A Bibliographic Review. In *Ideas and Methods in Quantum and Statistical Physics (Oslo, 1988)*; Cambridge University Press: Cambridge, UK, 1992; pp. 370–389.
13. Nelson, E. Construction of quantum fields from Markoff fields. *J. Funct. Anal.* **1973**, *12*, 97–112.
14. Kahn, J.; Kalai, G.; Linial, N. The Influence of Variables on Boolean Functions. In Proceedings of the IEEE Computer Society SFCS '88 29th Annual Symposium on Foundations of Computer Science, White Plains, NY, USA, 24–26 October 1988; pp. 68–80.
15. O'Donnell, R. *Analysis of Boolean Functions*; Cambridge University Press: Cambridge, UK, 2014.
16. Bonami, A. Étude des coefficients de Fourier des fonctions de $L^p(G)$. *Annales de l'institut Fourier* **1970**, *20*, 335–402. (In French)
17. Beckner, W. Inequalities in Fourier analysis. *Ann. Math.* **1975**, *102*, 159–182.
18. Gross, L. Hypercontractivity and logarithmic Sobolev inequalities for the Clifford-Dirichlet form. *Duke Math. J.* **1975**, *42*, 383–396.
19. Ahlswede, R.; Gács, P. Spreading of sets in product spaces and hypercontraction of the Markov operator. *Ann. Probab.* **1976**, *4*, 925–939.
20. Mossel, E.; Oleszkiewicz, K.; Sen, A. On reverse hypercontractivity. *Geom. Funct. Anal.* **2013**, *23*, 1062–1097.
21. Nair, C. (Chinese University of Hong Kong, Hong Kong, China); Kamath, S. (PDT Partners, Princeton, NJ, USA). Personal communication, 2016.
22. Alemi, A.A.; Fischer, I.; Dillion, J.V.; Murphy, K. Deep variational information bottleneck. *arXiv* **2017**, arXiv:1612.0041.
23. Achille, A.; Soatto, S. Information Dropout: Learning Optimal Representations Through Noisy Computation. *arXiv* **2016**, arXiv:1611.01353.

24. Nair, C. An extremal inequality related to hypercontractivity of Gaussian random variables. In Proceedings of the Information Theory and Applications Workshop, San Diego, CA, USA, 9–14 February 2014.
25. Gao, W.; Kannan, S.; Oh, S.; Viswanath, P. Conditional Dependence via Shannon Capacity: Axioms, Estimators and Applications. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 2780–2789.
26. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Washington, DC, USA, 28 June–2 July 2011; pp. 689–696.
27. Srivastava, N.; Salakhutdinov, R.R. Multimodal learning with deep boltzmann machines. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 2222–2230.
28. Andrew, G.; Arora, R.; Bilmes, J.; Livescu, K. Deep canonical correlation analysis. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 1247–1255.
29. Makur, A.; Zheng, L. Linear Bounds between Contraction Coefficients for f-Divergences. *arXiv* **2017**, arXiv:1510.01844.
30. Witsenhausen, H.S. On Sequences of Pairs of Dependent Random Variables. *SIAM J. Appl. Math.* **1975**, *28*, 100–113.
31. Bell, C. Mutual information and maximal correlation as measures of dependence. *Ann. Math. Stat.* **1962**, *33*, 587–595.
32. Nair, C. Equivalent Formulations of Hypercontractivity Using Information Measures. In Proceedings of the 2014 International Zurich Seminar on Communications, Zurich, Switzerland, 26–28 February 2014.
33. Chechik, G.; Globerson, A.; Tishby, N.; Weiss, Y. Information Bottleneck for Gaussian Variables. *J. Mach. Learn. Res.* **2005**, *6*, 165–188.
34. Michaeli, T.; Wang, W.; Livescu, K. Nonparametric Canonical Correlation Analysis. In Proceedings of the ICML'16 33rd International Conference on International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; Volume 48, pp. 1967–1976.
35. Simon, N.; Tibshirani, R. Comment on "Detecting Novel Associations In Large Data Sets" by Reshef Et Al, Science 16 December 2011. *arXiv* **2014**, arXiv:stat.ME/1401.7645.
36. Gorfine, M.; Heller, R.; Heller, Y. Comment on Detecting Novel Associations in Large Data Sets. Available online: http://emotion.technion.ac.il/~gorfinm/filesscience6.pdf (accessed on 30 October 2017).
37. Hastings, C.; Mosteller, F.; Tukey, J.W.; Winsor, C.P. Low Moments for Small Samples: A Comparative Study of Order Statistics. *Ann. Math. Stat.* **1947**, *18*, 413–426.
38. McBean, E.A.; Rovers, F. *Statistical Procedures for Analysis of Environmental Monitoring Data and Assessment*; Prentice-Hall: Upper Saddle River, NJ, USA, 1998.
39. Rustum, R.; Adeloye, A.J. Replacing outliers and missing values from activated sludge data using Kohonen self-organizing map. *J. Environ. Eng.* **2007**, *133*, 909–916.
40. Kumar, G. Binary Rényi Correlation: A Simpler Proof of Witsenhausen's Result and a Tight Lower Bound. Available online: http://www.gowthamiitm.com/research/Witsenhausen_simpleproof.pdf (accessed on 30 October 2017).