Conditional Dependence via Shannon Capacity: Axioms, Estimators and Applications

Weihao Gao WGAO9@ILLINOIS.EDU
CSL and Dept. of ECE, University of Illinois, Urbana-Champaign, USA

Sreeram Kannan KSREERAM@UW.EDU
Dept. of EE, University of Washington, Seattle, USA

Sewoong Oh SWOH@ILLINOIS.EDU
CSL and Dept. of IESE, University of Illinois, Urbana-Champaign, USA

Pramod Viswanath PRAMODV@ILLINOIS.EDU
CSL and Dept. of ECE, University of Illinois, Urbana-Champaign, USA

Abstract

We consider axiomatically the problem of estimating the strength of a conditional dependence relationship $P_{Y|X}$ from a random variables X to a random variable Y. This has applications in determining the strength of a known causal relationship, where the strength depends only on the conditional distribution of the effect given the cause (and not on the driving distribution of the cause). Shannon capacity, appropriately regularized, emerges as a natural measure under these axioms. We examine the problem of calculating Shannon capacity from the observed samples and propose a novel fixed-k nearest neighbor estimator, and demonstrate its consistency. Finally, we demonstrate an application to single-cell flow-cytometry, where the proposed estimators significantly reduce sample complexity.

1. Introduction

The axiomatic study of dependence measures on joint distributions between two random variables X and Y has a long history in statistics (Shannon, 1948; Rényi, 1959; Csiszár, 2008). In this paper, we study the relatively unexplored terrain of measures that depend only on the conditional distribution $P_{Y|X}$. We are motivated to study conditional dependence measures from a problem in causal strength estimation. Causal learning is a basic problem in many areas of scientific learning, where one wants to uncover the cause-effect relationship usually using interventions or sometimes directly from observational data (Pearl, 2009; Richardson & Evans, 2015; Mooij et al., 2015).

In this paper, we are interested in an even simpler question: given a causal relationship, how does one measure the strength of the relationship. This problem arises in many contexts, for example, one may know causal genetic pathways but only a subset of these maybe active in a particular tissue or organ - therefore, deducing how much influence each causal link exerts becomes necessary.

We focus on a simple model: consider a pair of random variables (X, Y) with *known* causal direction $X \to Y$, and suppose that there are no confounders - we are interested in *quantifying* the causal influence X has on Y. We denote the causal influence quantity by $\mathcal{C}(X, Y)$. There are two philosophically distinct ways to model the quantity: the first one is *factual influence*, i.e., how much influence does X exert on Y under the current probability of the cause X. The second possible way, which one can term as *potential influence* measures how much influence X can potentially exert on Y - without cognizance to the present distribution of the cause. For example, consider a (hypothetical) city which has very few smokers, but smoking inevitably leads to lung-cancer. In such a city, the factual influence of smoking on lung-cancer will be small but the potential influence is very high. Depending on the setting, one may prefer the former or the latter. In this paper, we are interested in the *potential influence* of a cause on its effect.

We want $\mathcal{C}(X, Y)$ to be invariant to scaling and one-one transformations of the variables X, Y. This naturally suggests information theoretic metrics as plausible choices of

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

$\mathcal{C}(X,Y)$, starting with the mutual information $I(X;Y) = D(P_{XY}||P_XP_Y)$, at least in the case of factual influence. This measures the information through the channel from $X \to Y$ as given by the prior P_X. Observe that this metric is *symmetric* with respect to the directions $X \to Y$ and $Y \to X$; this property is not always desirable. In fact, this measure is taken as a starting point to develop an axiomatic approach to studying causal strength on general graphs in (Janzing et al., 2013).

In a recent work (Krishnaswamy et al., 2014), potential causal influence is posited as a relevant metric to spot "trends" in gene pathways. In the particular application considered there, rare biological states of gene X in a given data may nevertheless correspond to important biological states (or become common under different biological conditions), and therefore it is important to have causal measures that are not sensitive to the cause distribution but only depend on the relationship between the cause and the effect. To quantify the potential influence of those rare X, the following approach is proposed. Replace the observed distribution P_X by a *uniform* distribution U_X and calculate the mutual information under the joint distribution $U_XP_{Y|X}$. The resulting causal strength quantification is $\mathcal{C}(X,Y) = D(U_XP_{Y|X}||P_UP_Y)$, where P_Y represents the distribution at the output of a channel $P_{Y|X}$ with input given by U_X. We call this quantification as Uniform Mutual Information (UMI) and pronounced "you-me". A key challenge is to compute this quantity from i.i.d. samples in a statistical efficient manner, especially when the channel output is continuous valued (and potentially in high dimensions). This is the first focus point of this paper.

UMI is not invariant under bijective transformations (since a uniform distribution on X is different from a uniform distribution on X^3) and is also sensitive to the estimated support size of X. Even more fundamentally, it is unclear why one would prefer the uniform prior to measure potential influence through the channel $P_{Y|X}$. Based on natural axioms of data processing and additivity, we motivate an alternative measure of causal strength: the *largest amount of information* that can be sent through the channel, namely the *Shannon capacity*. Formally $\mathcal{C}(X,Y) = \max_{Q_X} D(Q_XP_{Y|X}||Q_XP_Y)$, where P_Y represents the distribution at the output of a channel $P_{Y|X}$ with input given by Q_X. We refer to such a quantification as Capacitated Mutual Information (CMI) and pronounced "see-me". A key challenge is to compute this quantity from i.i.d. samples in a statistical efficient manner, especially when the channel output is continuous valued (and potentially in high dimensions). This is the second focus point of this paper. We make the following **main contributions** in this paper.

- *UMI Estimation*: We construct a novel estimator to compute UMI from data sampled i.i.d. from a distribution P_{XY}. The estimator brings together ideas from three disparate threads in statistical estimation theory: nearest-neighbor methods, a correlation boosting idea in the estimation of (standard) mutual information from samples (Kraskov et al., 2004), and importance sampling. The estimator has only a *single* hyper parameter (the number of nearest-neighbors considered, set to 4 or 5 in practice), uses an off-the-shelf kernel density estimator of only P_X, and has strong connections to the entropy estimator of (Kozachenko & Leonenko, 1987). Our main technical result is to show that the estimator is consistent (in probability) supposing that the Radon-Nikodym derivative $\frac{dP_U}{dP_X}$ is uniformly bounded over the support. In simulations, the estimator has very strong performance in terms of sample complexity (compared to a baseline of the partition-based estimator in (Moddemeijer, 1989)).

- *CMI Estimation*: We build upon the estimator derived for UMI and construct an optimization problem that mimics the optimization problem inherent in computing the capacity directly from the conditional probability distribution of the channel. Our main technical result is to show the consistency of this estimator, supposing that the Radon-Nikodym derivative $\frac{dP_Q}{dP_X}$ is uniformly bounded over the support, where P_Q is the optimizing input to the channel. Simulation results show strong empirical performance, compared to a baseline of a partition-based method followed by discrete optimization.

- *Application to gene pathway influence*: In (Krishnaswamy et al., 2014), considered an important result in single-cell flow-cytometry data analysis, a causal strength metric (termed DREMI) is proposed for measuring the causal influence of a gene – this estimator is a specific way of implementing UMI along with a "channel amplification" step, and DREMI was successfully used to spot gene-pathway trends. We show that our proposed CMI and UMI estimators also exhibit the same performance as DREMI when supplied with the full dataset, while at the same time, having significantly smaller sample complexity for the same performance.

2. An Axiomatic Approach

We formally model an influence measure on conditional probability distributions, by postulating five natural axioms. Let X be drawn from an alphabet $\mathcal{X}$, and Y from an alphabet $\mathcal{Y}$. Let the probability distribution of Y given X be given as $P_{Y|X}$. Let $\mathcal{P}$ be a family of conditional distributions; usually we will consider the case when $\mathcal{P}$ is the set of all possible conditional distributions. Then the influ-

ence measure $\mathcal{C}$ is a function of the conditional distribution to non-negative real numbers: $\mathcal{C} : \mathcal{P}(\mathcal{Y}|\mathcal{X}) \to \mathbb{R}^+$. We postulate that the function $\mathcal{C}$ satisfies five axioms on $\mathcal{P}$, and show that CMI satisfies all five axioms:

0. **Independence**: The measure $\mathcal{C}(P_{Y|X}) = 0$ if and only if $P_{Y=y|X=x}$ only depends on y.

1. **Data Processing**: If $P_{Z=z|X=x} = \sum_{y\in\mathcal{Y}} P_{Z=z|Y=y}P_{Y=y|X=x}$, i.e., $X \to Y \to Z$ be a processing chain, then the natural data processing inequalities should hold: (a) $\mathcal{C}(P_{Y|X}) \geq \mathcal{C}(P_{Z|X})$; and (b) $\mathcal{C}(P_{Z|Y}) \geq \mathcal{C}(P_{Z|X})$.

2. **Additivity**: For a parallel channel $P_{Y_1,Y_2|X_1,X_2} := P_{Y_1|X_1}P_{Y_2|X_2}$, we need $\mathcal{C}(P_{Y_1,Y_2|X_1,X_2}) = \mathcal{C}(P_{Y_1|X_1}) + \mathcal{C}(P_{Y_2|X_2})$.

3. **Monotonicity**: A causal relationship is strong if many possible values of P_Y are achievable by varying the input probability distribution P_X. Thus if we consider $P_{Y|X}$ as a map from the probability simplex in X to the probability simplex in Y, the larger the range of this map, the stronger should be the causal strength.

   (a) $\mathcal{C}$ should only depend on the range of the map, Range$(P_{Y|X})$, the convex hull of the output distributions $P_{Y|X=x}$.

   (b) $\mathcal{C}$ should be a monotonic function of the range of the map. If $P_{Y|X}$ and $Q_{Y|X}$ are such that, Range$(P_{Y|X}) \subseteq$ Range$(Q_{Y|X})$ then: $\mathcal{C}(P_{Y|X}) \leq \mathcal{C}(Q_{Y|X})$.

4. **Maximum value**: The maximum value over all possible conditional distributions for a particular output alphabet $\mathcal{Y}$ should be achieved exactly when the relationship is fully causal, i.e., each $Y = y$ can be achieved by setting $X = x$ for some x.

We begin our exploration of appropriate influence measures with the alphabets for X and Y being discrete. Let $I(P_{XY}) := D(P_{XY}||P_XP_Y)$ denote the mutual information with respect to the joint distribution P_{XY}. Since we are looking at *potential* influence measures, Shannon capacity, defined as the maximum over input probability distributions of the mutual information, is a natural choice:

$$\text{CMI}(P_{Y|X}) := \max_{P_X} I(P_X P_{Y|X}). \tag{1}$$

Our first claim is that

CMI *satisfies all the axioms of causal influence.*

The proof (omitted in this conference version) is available in the full version (Gao et al., 2016a).

Axiomatic View of UMI : Now consider an alternative metric: Uniform Mutual Information (UMI) which is defined as the mutual information with uniform input distribution,

$$\text{UMI}(P_{Y|X}) := I(U_X P_{Y|X}), \tag{2}$$

where U_X is the uniform distribution on $\mathcal{X}$. This estimator is motivated by the recent work in (Krishnaswamy et al., 2014). We investigate how this estimator fares in terms of the proposed axioms.

- UMI clearly satisfies Axiom 0. It also satisfies Axioms $1a$. Data-processing inequality for mutual information on the joint distribution $U_X P_{Y|X} P_{Z|Y}$ implies that $I(U_X P_{Y|X}) \geq I(U_X P_{Z|X})$, which is the same as UMI$(P_{Y|X}) \geq$ UMI$(P_{Z|X})$. Thus $I(U_Y P_{Z|Y}) \geq I(U_X P_{Z|X})$.

- UMI however does not satisfy Axiom $1b$ in general. However, if the transition matrices $P_{Y|X}$ and $P_{Z|Y}$ are both doubly stochastic, then a straightforward calculation shows that UMI satisfies Axiom $1b$ too.

- UMI satisfies Axiom 2 since the uniform distribution on X_1, X_2 naturally factors as $U_{X_1,X_2} = U_{X_1}U_{X_2}$ and we have UMI$(P_{Y_1,Y_2|X_1,X_2})$

$$\begin{aligned} &= I(U_{X_1,X_2} P_{Y_1,Y_2|X_1,X_2}) && (3)\\ &= I(U_{X_1}U_{X_2} P_{Y_1|X_1} P_{Y_2|X_2}) && (4)\\ &= \text{UMI}(P_{Y_1|X_1}) + \text{UMI}(P_{Y_2|X_2}). && (5) \end{aligned}$$

- UMI does not satisfy Axiom $3a$ since multiple repeated values of $P_{Y|X=x}$ does not alter the convex hull but alters the UMI value.

- Interestingly, UMI does satisfy Axiom 4 for the same reason as CMI.

2.1. Real-valued alphabets

For real-valued X, the Shannon capacity is not finite without additional regularizations. This is also true of other measures of relation such as the Renyi correlation (Rényi, 1959), and in each case the measure is studied in the context of some form penalty term. Typically this is done via a cost constraint on the real-valued input parameters. In this context, one possibility is to consider the following norm-constrained optimization to ensure the causal effect is finite valued:

$$\text{CMI}(P_{Y|X}, a) := \max_{P_X : \mathbb{E}\|X\|_2^2 \leq a} I(P_X P_{Y|X}). \tag{6}$$

In practice, a is chosen from the empirical moments of X from samples: $a := \frac{1}{N}\sum_{i=1}^N \|X_i\|_2^2$ for samples $X_1, \ldots, X_N$. This regularization turns to be the so-called power constraint on the input, common in treatments of additive noise communication channels.

3. Estimators

Although the definition of UMI and CMI seamlessly applies to both discrete and continuous random variables, the estimation becomes relatively straightforward when both $\mathcal{X}$ and $\mathcal{Y}$ are discrete; the estimation of the conditional distribution $p_{Y|X}$ and the the computation of UMI and CMI can be separated in a straightforward manner. For this reason and also due to the motivation in genomic biology that we study, we focus on the more challenging regime $\mathcal{Y}$ is continuous. Due to certain subtleties in the estimation process, we provide separate estimators each customized for each case of discrete and continuous $\mathcal{X}$, respectively.

3.1. Uniform Mutual Information

The idea of applying UMI to infer the strength of conditional dependence was first proposed in (Krishnaswamy et al., 2014). Off-the-shelf 2-dimensional kernel density estimators (KDE) are used to first estimate the joint distribution P_{XY} from given samples. Subsequently, the channel $P_{Y|X}$ is computed directly from the joint, and then UMI is computed via either numerical integration or sampling from U_X and $P_{Y|X}$. This approach suffers from known drawbacks of KDE, such as sensitivity to the choice of the bandwidth and increased bias in higher dimensional X and Y. However, a more critical challenge in using KDE to estimate the joint at all points (and not just at samples) is the *overkill* nature: we only need to compute a single functional (UMI) of the joint distribution, which could in principle be computed more efficiently directly from the samples. It is not at all clear how to *directly* estimate UMI, where X is changed to uniform.

Perhaps surprisingly, we bring together ideas from three topics in statistical estimation to introduce novel estimators that are also provably convergent. Our estimator is based on (a) k-nearest neighbor estimators, e.g. (Kozachenko & Leonenko, 1987); (b) the correlation boosting idea of the estimator from (Kraskov et al., 2004)–which is widely adopted in practice (Khan et al., 2007); and (c) the importance sampling techniques to adjust for the uniform prior for UMI. We explain each idea below.

Consider a simpler task of computing the mutual information from samples; several approaches exist for this estimation: (Paninski, 2003; Kraskov et al., 2004; Wang et al., 2009; Pál et al., 2010; Sricharan et al., 2010; Póczos et al., 2012; Gao et al., 2014; 2015; Kandasamy et al., 2015). Note that three applications of the entropy estimator, such as those from (Beirlant et al., 1997), gives an estimate of the mutual information, i.e. $\widehat{I}(X;Y) = \widehat{H}(X) + \widehat{H}(Y) - \widehat{H}(X,Y)$. Each entropy term can be computed using, for example, a KDE based approach, which suffers from the same challenges, as in UMI. Alternatively, to bypass estimating P_{XY} at every point, the differential entropy estimation can be done via k nearest neighbor (kNN) methods (pioneering work in (Kozachenko & Leonenko, 1987)). This KL entropy estimator provides the first step in designing the UMI estimator. However, taking the route of estimating the mutual information via estimating the three differential entropies (two marginals and one joint), it is entirely unclear how to estimate two of these quantities (differential entropy of Y and that of (U,Y)) directly from samples.

Perhaps surprisingly, an innovative approach undertaken in (Kraskov et al., 2004) to improve upon three applications of KL estimators provides a solution. The KSG estimator of (Kraskov et al., 2004) is based on kNN distance $\rho_{k,i}$ defined as the distance to the k-th nearest neighbor from (X_i, Y_i) in ℓ_∞ distance, i.e. $\rho_{k,i} = \max\{\|X_{j_k} - X_i\|_\infty, \|Y_{j_k} - Y_i\|_\infty\}$ where (X_{j_k}, Y_{j_k}) is the k-th nearest neighbor to (X_i, Y_i). Precisely, the KSG estimator is $\widehat{I}(X;Y) =$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\psi(k) + \psi(N) - \psi(n_{x,i}) - \psi(n_{y,i})\right), \tag{7}$$

where $\psi(x)$ is the digamma function, $\psi(x) = \Gamma'(x)/\Gamma(x)$ (for large x, $\psi(x) \approx \log x - 1/(2x)$), and the kNN statistics $n_{x,i}$ and $n_{y,i}$ are defined as

$$n_{x,i} \equiv \sum_{j\neq i}\mathbb{I}\{\|X_j - X_i\|_\infty < \rho_{k,i}\}, \text{ and} \tag{8}$$

$$n_{y,i} \equiv \sum_{j\neq i}\mathbb{I}\{\|Y_j - Y_i\|_\infty < \rho_{k,i}\}. \tag{9}$$

Note that the number of nearest neighbors in X and Y are computed with respect to $\rho_{k,i}$ in the joint space (X,Y). This innovative idea, not only gives a simple estimator, but also has an advantage of canceling correlations in three entropy estimates, giving an improved performance. However, despite its popularity in practice due to its simplicity, no convergence result is known.

Inspired by the innovative mutual information estimator in (7), we combine importance sampling techniques to adjust for the uniform prior for UMI, and propose a novel estimator. On top of the provable convergence, our estimator has only one hyper-parameter k (besides the choice of bandwidth h_N for estimating the marginal distribution P_X which is a significantly simpler task compared to estimating the joint), which is the number of nearest neighbors to consider; in practice k is set to a small integer such as 4 or 5.

Continuous $\mathcal{X}$. We propose a novel UMI estimator based on the Kraskov mutual information estimator. For a conditional probability density $f_{Y|X}$, we want to compute the uniform mutual information from N i.i.d. samples $(X_1, Y_1), \ldots, (X_N, Y_N)$ that are generated from $f_{Y|X} f_X$

for some prior on X. Our UMI estimator is based on k nearest neighbor (kNN) statistics. Given a choice of $k \in \mathbb{Z}^+$ and N samples,

$$\widehat{\mathrm{UMI}} \equiv \frac{1}{N}\sum_{i=1}^{N} w_i \left(\psi(k) + \log \frac{N c_{d_x} c_{d_y}}{c_{d_x+d_y}\, n_{x,i}\, n_{y,i}}\right), \quad (10)$$

where $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$, $c_d = \pi^{\frac{d}{2}}/\Gamma(\frac{d}{2}+1)$ is the volume of d-dimensional unit ball, and w_i is the *self-normalized* importance sampling estimate (Cornuet et al., 2012) of $\frac{u(X_i)}{f(X_i)}$:

$$w_i \equiv \frac{N/\tilde{f}(X_i)}{\sum_{j=1}^{N}\left(1/\tilde{f}(X_j)\right)}, \quad (11)$$

where $\tilde{f} : \mathcal{X} \to \mathbb{R}$ is the estimate of $f_X(x)$. We use the standard kernel density estimator with a bandwidth h_N:

$$\tilde{f}(x) \equiv \frac{1}{N h_N^{d_x}} \sum_{i=1}^{N} K\Big(\frac{X_i - x}{h_N}\Big). \quad (12)$$

We define the kNN statistics $n_{x,i}$ and $n_{y,i}$ as follows. For each sample (X_i, Y_i), calculate the Euclidean distance $\rho_{k,i}$ (as opposed to the ℓ_∞ distance proposed by Kraskov et al. (2004)) to the k-th nearest neighbor. This determines the (random) number of samples within $\rho_{k,i}$ in $\mathcal{X}$: first $n_{x,i}$ is defined as the same as in (8), but with Euclidean distance; second we have a *weighted* number of samples within $\rho_{k,i}$ in $\mathcal{Y}$ as

$$n_{y,i} \equiv \sum_{j\neq i} w_j \mathbb{I}\{\|Y_j - Y_i\| < \rho_{k,i}\}. \quad (13)$$

Compared to (7), we first exchange log function for the digamma functions of N, $n_{x,i}$, and $n_{y,i}$. This step (especially for $n_{x,i}$, and $n_{y,i}$) is crucial for proving convergence. We use ideas from importance sampling and introduce new variables w_i's that capture the correction for the mismatch in the prior. The constants c_{d_x}, c_{d_y}, and $c_{d_x+d_y}$ correct for the volume measured in ℓ_2.

Discrete $\mathcal{X}$. Similarly, for a discrete random variable X, the joint probability density function is denoted by $f(x,y) = p_X(x) f_{Y|X}(y|x)$. We propose a UMI estimator, and overload the same notation for this discrete case.

$$\widehat{\mathrm{UMI}} \equiv \frac{1}{N}\sum_{i=1}^{N} w_{X_i} \left(\psi(k) + \log \frac{N}{n_{X_i}\, n_{y_i}}\right), \quad (14)$$

where $n_{X_i} = |\{j \in [N] : j \neq i, X_j = X_i\}|$ is the number of samples j such that $X_j = X_i$, w_{X_i} is the self-normalizing estimate of $1/(|\mathcal{X}| p_X(X_i))$ defined as

$$w_x \equiv \frac{N}{|\mathcal{X}| n_x}, \quad (15)$$

and $n_{y,i}$ is the weighted kNN statistics defined as follows. For each sample (X_i, Y_i), let the distance to the k-th nearest neighbor be $\rho_{k,i}$, where those samples that have the same X value as X_i is considered and the Euclidean distance is measured in $\mathcal{Y}$. We define the *weighted* number of samples within $\rho_{k,i}$ in $\mathcal{Y}$ as

$$n_{y,i} \equiv \sum_{j\neq i} w_{X_j} \mathbb{I}\{\|Y_j - Y_i\| < \rho_{k,i}\}. \quad (16)$$

3.2. Capacitated Mutual Information

Given standard estimators for mutual information and entropy, it is not at all clear how to *directly* estimate CMI where f_X is changed to the (unknown) optimal input distribution. However, combining the mutual information estimator in (7) with importance sampling techniques, we design a novel estimator as a solution to an optimize over the space of the weights. Our estimator has only one hyperparameter k, the number of nearest neighbors to consider.

Continuous $\mathcal{X}$. For a conditional distribution $f_{Y|X}$, we compute an estimate of CMI from i.i.d. samples $(X_1, Y_1), \ldots, (X_N, Y_N)$ generated from $f_{Y|X} f_X$ for some prior on X. We introduce a novel CMI estimator that is based on our UMI estimator. Given a choice of $k \in \mathbb{Z}^+$ and N samples, the estimated CMI is the solution of the following constrained optimization:

$$\widehat{\mathrm{CMI}} = \max_{w \in T_{a,N}} \frac{1}{N}\sum_{i=1}^{N} w_i \left(\psi(k) + \log\Big(\frac{N c_{d_x} c_{d_y}}{c_{d_x+d_y} n_{x,i} n_{y,i}}\Big)\right),$$

where d_x, d_y, $n_{x,i}$, $n_{y,i}$ and c_d are defined in the same as in (10). We optimize over $w_1, \ldots, w_N$ under the second moment constraint, i.e. $T_{a,N} = \{w \in \mathbb{R}^N | w_i \geq 0, \forall i \in [N], (1/N)\sum_{i=1}^{N} w_i = 1, (1/N)\sum_{i=1}^{N} w_i \|X_i\|^2 \leq a^2\}$. Observe that no KDE of P_X is needed for CMI estimation, making it particularly simple and robust.

Discrete $\mathcal{X}$. Similarly, we define the CMI estimate $\widehat{\mathrm{CMI}}$ as the solution of the following constrained optimization:

$$\widehat{\mathrm{CMI}} = \max_{w \in T_\Delta} \frac{1}{N}\sum_{i=1}^{N} w_{X_i} \left(\psi(k) + \log\Big(\frac{N}{n_{x,i} n_{y,i}}\Big)\right)$$

where $n_{x,i}$ and $n_{y,i}$ are defined in (14). T_Δ is the set of quantized version of an interval $[C_1, C_2]$ with step size Δ, i.e. $T_\Delta = \big\{w \in \{C_1 + m_i\Delta\}^{|\mathcal{X}|} \big| (1/N)\sum_{x=1}^{|\mathcal{X}|} w_x \in [1 - |\mathcal{X}|\Delta, 1 + |\mathcal{X}|\Delta]$, and $m_i \in \{0, 1, \ldots, \lceil (C_2 - C_1)/C_1 \rceil\}$ for all $i\big\}$. Such a quantization is crucial in proving consistence in Theorem 2.

4. Convergence Guarantees

We show both UMI and CMI estimators we propose are consistent under typical assumptions on the distribution.

Uniform Mutual Information: As our estimators use the off-the-shelf kernel density estimator of P_X (Devroye & Penrod, 1984; Sheather & Jones, 1991) and also the ideas from the nearest-neighbor methods (Kozachenko & Leonenko, 1987), we make assumptions on the conditional density $f_{Y|X}$ that are typical in these literature. One extra assumption we make for UMI is that the Radon-Nikodym derivative $\frac{dP_U}{dP_X}$ is uniformly bounded over the support. This is necessary for controlling the importance-sampling estimates of w_i's. We refer to the Assumption 1 provided in the longer version of this paper (Gao et al., 2016a) for a precise description.

Theorem 1. *Under the Assumption 1 in (Gao et al., 2016a), the UMI estimator converges to the true value in probability, i.e. for all $\varepsilon > 0$ and all $\delta > 0$,*

$$\lim_{N\to\infty} \mathbb{P}\left(\left| \widehat{\mathrm{UMI}} - \mathrm{UMI}(f_{Y|X}) \right| > \varepsilon \right) = 0 \,, \qquad (17)$$

if $k > \max\{d_y/d_x, d_x/d_y\}$ for continuous X and $(\log N)^{(1+\delta)d_y} < k < \sqrt{N}/(5\log N)$ for discrete X.

In practice, we regularize the kNN distance $\rho_{k,i}$ in case it is much smaller than the expected distance of order $N^{-1/(d_x+d_y)}$. For continuous $\mathcal{X}$, we require k to be larger than the ratio of the dimensions, which is a finite constant. For discrete $\mathcal{X}$, however, the effective dimension of $\mathcal{X}$ is zero, which makes the ratio d_y/d_x unbounded. Hence, for concentration of measure to hold, we need k^{1/d_y} scaling at least logarithmically in the number of samples N.

Capacitated Mutual Information: We make analogous assumptions which are described precisely in Assumption 2 provided in the longer version of this paper (Gao et al., 2016a). The following theorem establishes consistency of our estimator when $\mathcal{X}$ is discrete and we quantize $\mathcal{Y}$. Our analysis requires uniform convergence over all possible choices of the weights w, making the quantization step inevitable; improvements on this technical condition are natural future steps.

Theorem 2. *Under the Assumption 2 in (Gao et al., 2016a), the CMI estimator converges in probability to the true value up to the resolution of the quantization, i.e. if $k > (\log N)^{(1+\delta)d_y}$ for some $\delta > 0$, and $k < \sqrt{N}/(5\log N)$, for all $\varepsilon > 0$ and $\Delta > 0$, there exists $s(\Delta) = O(\Delta)$ such that*

$$\lim_{N\to\infty} \mathbb{P}\left(\left| \widehat{\mathrm{CMI}} - \mathrm{CMI}(f_{Y|X}) \right| > \varepsilon + s(\Delta) \right) = 0.$$

5. Numerical Experiments

5.1. Gene Causal Strength from Single Cell Data

We briefly describe the setup of (Krishnaswamy et al., 2014) to motivate our numerical experiments. Consider

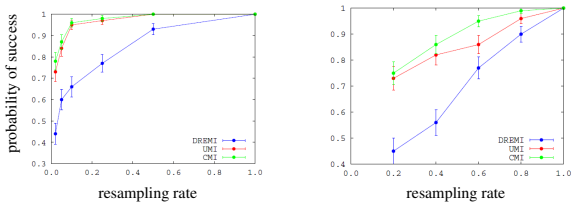

Figure 1. CMI and UMI estimators significantly improve over DREMI in capturing the biological trend in flow-cytometry data: the figures above refer to the same setting as Figure 6 of (Krishnaswamy et al., 2014).

a simple genetic pathway: a cascade of genes having expression values X, Y, Z which interact linearly, i.e., $X \to Y \to Z$. A key question of interest in this case is how the signaling in the pathway varies in different conditions of intervention. Let T denote the time after the intervention (for example, after giving a certain drug). Then we may want to compare the strength of the causal relationship between two genes at different times after the intervention. In the experiments, usually samples are taken at very few time points, so T has very small cardinality (for example, before the drug, 10 minutes after the drug and 50 minutes after the drug), but at each given time point, many cells are interrogated so we get samples from the distribution $P_{X,Y,Z;T=t} = P(Y|X;T=t)P(Z|Y;T=t)$. For each value of $T = t$, we observe N_t i.i.d. samples (X_i, Y_i, Z_i), for $i = 1, 2, ..., N_t$ sampled from $P_{X,Y,Z;T=t}$. These samples are obtained using a technique called single-cell mass flow cytometry, see (Krishnaswamy et al., 2014) for details. We are interested in obtaining a causal measure $\mathcal{C}(X \to Y; T=t) = \mathcal{C}(P(Y|X;T=t))$ and another measure $\mathcal{C}(Y \to Z; T=t) = \mathcal{C}(P(Z|Y;T=t))$ for each time point t. This measure serves as a high level summary of how signaling proceeds in the cascade as a function of time, and lets one compare the strengths of a given causal relationship at different points after intervention.

If the drug indeed activates the causal pathway, one may expect the causal relationship to follow a certain *trend*, i.e., at earlier t, the strength of $\mathcal{C}(X \to Y; T=t)$ will be high and at a later value of t, the strength of $\mathcal{C}(Y \to Z; T=t)$ will be high before the effect of the drug wears off, at which time we expect all the relationships to fall back to its low nominal value. Such an analysis is conducted in (Krishnaswamy et al., 2014) where the causal strength function $\mathcal{C}$ is evaluated via the so-called DREMI estimator (essentially a version of UMI estimation with a "channel amplification" step and careful choice of hyper parameters therein – no theoretical properties of this estimator were evaluated). In that paper, it is shown that, for two example pathways, DREMI recovers the correct trend, i.e., it correctly identifies the time at which each causal rela-

tionship is expected to peak as per prior biological knowledge. This demonstrates the utility of DREMI for causal strength inference in gene networks (see Figure 6 of (Krishnaswamy et al., 2014)). The authors there also demonstrate that other metrics which depend on the whole joint distribution, such as mutual information, maximal information coefficient, and correlation do not capture the trend. As an aside, we note that a somewhat different set of "trend spotting" estimators, primarily trying to find genes which demonstrate a monotonic trend over time from single-cell RNA-sequencing data, have been proposed very recently in (Mueller et al., 2015).

In this paper, we have studied influence measures axiomatically and proposed the UMI and CMI measures. It is natural to apply our estimators to *each time point* in the same setting as (Krishnaswamy et al., 2014) – and look to understand two distinct issues in our experiments with the flow-cytometry data. The first is whether the proposed quantities of UMI and CMI are able to capture the same biological trend as DREMI was able to. The second question relates to the sample complexity: how does the ability to recover the trend vary as a function of the sample complexity? To study this, we subsample the original data from (Krishnaswamy et al., 2014) multiple times (100 in the experiments) at each subsampling ratio and compute the fraction of times we recover the true biological trend. This is plotted in Figure 1. The figure demonstrates that when the whole dataset is made available, UMI and CMI are able to spot the trend correctly (just as DREMI does). When fewer samples are available, UMI uniformly dominates DREMI and, in turn, CMI uniformly dominates UMI in terms of capturing the biological trend as a function of number of samples available. We believe that this strong empirical evidence lends credence to our approach. For completeness, we note that the datasets represented in Figure 1 refer to regular T-cells (left figure) and T-cells exposed with an antigen (right figure), for which we expect different biological trends, but both of which are correctly captured by our metrics.

5.2. Synthetic data

We demonstrate the accuracy of the proposed UMI and CMI estimators on synthetic experiments. We generate N samples from P_{XY} where X is distributed as beta distribution $\mathrm{Beta}(1.5, 1.5)$ and $Y = X + N$, $N \sim \mathcal{N}(0, \sigma^2)$, independent of X. We present three results with varying $\sigma^2 \in \{0.09, 0.36, 1.0\}$. Figure 2 shows the estimate of UMI, averaged over 100 instances. This is compared to the ground truth and the state-of-the-art partition based estimators from (Moddemeijer, 1989). The ground truth has been computed via simulations with 8192 samples from the desired distribution $P_{Y|X}U_X$ using Kraskov's mutual information estimator (Kraskov et al., 2004). For CMI, we use exactly the same distribution P_{XY} as in UMI, but with varying $\sigma^2 \in \{0.36, 1.0, 2.25\}$, which is illustrated in Figure 3. Under the power constraint, the ground truth is given by $\frac{1}{2}\log(1 + \frac{\sigma_X^2}{\sigma_N^2}) = \frac{1}{2}\log(1 + 1/16\sigma^2)$. The proposed CMI estimator is compared against Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972) for computing discrete channel capacity, applied to quantized data. Both figures illustrate that the proposed estimators significantly improves over the state-of-the-art partition based methods, in terms of sample complexity.

6. Discussion

In this paper we have proposed novel information theoretic measures of potential influence of one variable on another, as well as provided novel estimators to compute the measures from i.i.d. samples. The technical innovation has been in proposing these estimators, by combining separate threads of ideas in statistics (including importance sampling and nearest-neighbor methods). The consistency proofs suggest that a similar analysis the very popular estimator of (traditional) mutual information in (Kraskov et al., 2004) can be conducted successfully; such work has been recently conducted in (Gao et al., 2016b). Several other issues in statistical estimation theory intersect with our current work and we discuss some of these topics below.

(a) The main technical results of this paper have been weak consistency of the proposed estimators. Proving stronger consistency guarantees and rates of convergence would be natural improvements, albeit challenging ones. Rates of convergence in the nearest-neighbor methods are barely known in the literature even for traditional information theoretic quantities: for instance, (Tsybakov & Van der Meulen, 1996) derives a $\sqrt{N}$ consistency for the single dimensional case of differential entropy estimation (under strong assumptions on the underlying pdf), leaving higher dimensional scenarios open, and which recently have been successfully addressed in (Gao et al., 2016b).

(b) There is a natural generalization of our estimators when the alphabet $\mathcal{Y}$ is high dimensional, using the kNN approach (just as in the differential entropy estimator of (Kozachenko & Leonenko, 1987) or in the mutual information estimator of (Kraskov et al., 2004)). However, very recent works (Gao et al., 2014; 2015; Lombardi & Pant, 2016) have shown that boundary biases common in high dimensional scenarios is much better handled using local parametric methods (as in (Loader et al., 1996; Hjort & Jones, 1996)). Adapting these approaches to the estimators for UMI and CMI is an interesting direction of future research.

(c) We have considered both the case of discrete and (single dimensional) continuous alphabet $\mathcal{X}$. The scenario of high dimensional $\mathcal{X}$ is significantly more challenging for

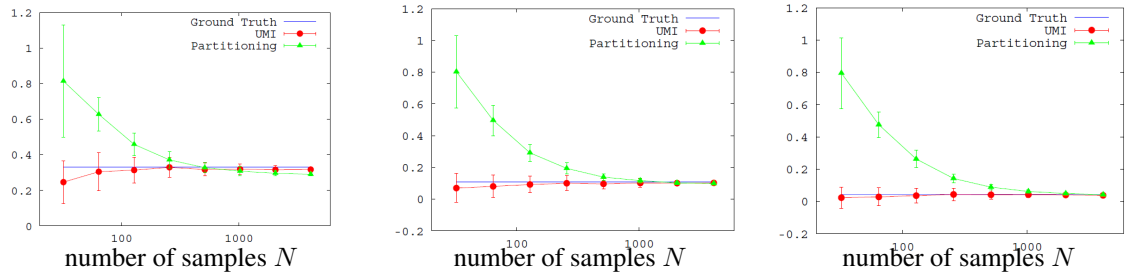

Figure 2. The proposed UMI estimator significantly outperforms partition based methods (Moddemeijer, 1989) in sample complexity. Additive Gaussian channels are used with varying variances σ^2: 0.09 (left), 0.36 (middle), and 1.0 (right).

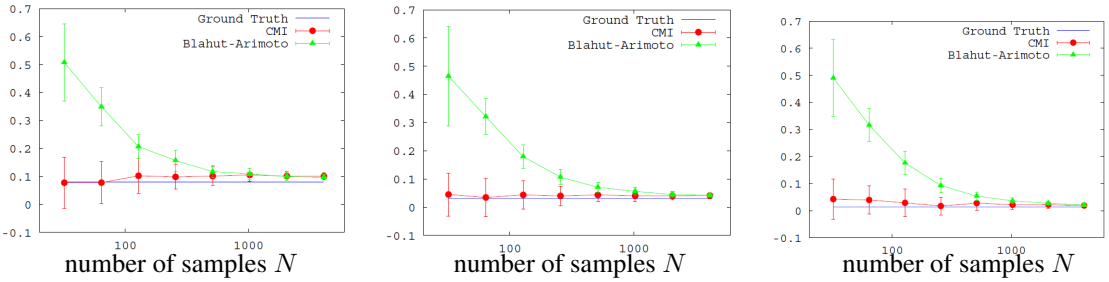

Figure 3. The proposed CMI estimator significantly outperforms partition based methods (Blahut, 1972; Arimoto, 1972) in sample complexity. Additive Gaussian channels are used with varying variances σ^2: 0.36 (left), 1.0 (middle), and 2.25 (right).

CMI estimation: this is because of the (vastly) expanded space of distributions over which the optimization can be performed. Also challenging is to consider application specific regularization of the inputs in this scenario.

(d) While the focus of this paper has been on quantifying potential causal influence, a related question involves testing the *direction* of causality for a pair of random variables. This is a widely studied topic with a long lineage (Pearl, 2009) but also of strong topical interest (Janzing et al., 2013; 2015; Mooij et al., 2015; Shajarisales et al., 2015). A natural inclination is to explore the efficacy of UMI and CMI measures to test for direction of causality – especially in the context of the benchmark data sets collected in (Mooij et al., 2015). Our results are as follows: UMI has a 45% probability to predict the correct direction. CMI gives 53% probability. Directly comparing the marginal entropy $H(X)$ and $H(Y)$ by the estimator in (Kozachenko & Leonenko, 1987) also only provides 45% accuracy. While in (Mooij et al., 2015), different entropy estimators (with appropriate hyper parameter choices) were applied to get an accuracy up to 60%-70%. Further research is needed to shed conclusive light, although we point out that the benchmark data sets in (Mooij et al., 2015) have substantial confounding factors that make causal direction hard to measure in the first place.

(e) The axiomatic derivation of potential causal influence naturally suggests CMI as an appropriate measure. We are also able to show (details are in a journal version (Gao et al., 2016a)) that a more general quantity (the so-called Rényi capacity, an appropriate maximum over all Rényi divergences) and which simplifies to the Shannon capacity with a specific parameter choice), also meets the axioms. It would be interesting to design estimators for the more general family of Rényi capacity measures, as would be to understand the role of additional axioms that would lead to uniqueness of Shannon capacity (in the same spirit as entropy being uniquely characterized by somewhat similar axioms (Csiszár, 2008)).

(f) Finally, a comment on the optimization problem in CMI estimation: the optimization problem involving the w_i's is not necessarily a concave program for a given sample realization, although one can show that for large enough sample size N the program is concave with high probability (indeed, in the limit of large sample size, this program converges to that of Shannon capacity computation involves maximizing mutual information, which is a concave function of the input probability distribution). Standard (stochastic) gradient decent is used in our experiments, and we did not face any disparity in convergent values over the set of synthetic experiments we conducted.

Acknowledgements

This work is supported in part by ARO W911NF1410220, NSF SaTC award CNS-1527754, NSF CISE award CCF-1553452 and a University of Washington startup grant.

References

Arimoto, Suguru. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *Information Theory, IEEE Transactions on*, 18(1):14–20, 1972.

Beirlant, Jan, Dudewicz, Edward J, Györfi, László, and Van der Meulen, Edward C. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

Blahut, Richard E. Computation of channel capacity and rate-distortion functions. *Information Theory, IEEE Transactions on*, 18(4):460–473, 1972.

Cornuet, Jean, Marin, Jean-Michel, Mira, Antonietta, and Robert, Christian P. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, 2012.

Csiszár, Imre. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.

Devroye, Luc and Penrod, Clark S. The consistency of automatic kernel density estimates. *The Annals of Statistics*, pp. 1231–1249, 1984.

Gao, Shuyang, Steeg, Greg Ver, and Galstyan, Aram. Efficient estimation of mutual information for strongly dependent variables. *arXiv preprint arXiv:1411.2003*, 2014.

Gao, Shuyang, Steeg, Greg Ver, and Galstyan, Aram. Estimating mutual information by local gaussian approximation. *arXiv preprint arXiv:1508.00536*, 2015.

Gao, Weihao, Kannan, Sreeram, Oh, Sewoong, and Viswanath, Pramod. Conditional dependence via shannon capacity: Axioms, estimators and applications. *arXiv preprint arXiv:1602.03476*, 2016a.

Gao, Weihao, Oh, Sewoong, and Viswanath, Pramod. Demystifying fixed k-nearest neighbor information estimators. *arXiv preprint arXiv:1604.03006*, 2016b.

Hjort, Nils Lid and Jones, MC. Locally parametric nonparametric density estimation. *The Annals of Statistics*, pp. 1619–1647, 1996.

Janzing, D., Steudel, B., Shajarisales, N., and Schölkopf, B. *Justifying Information-Geometric Causal Inference*, chapter 18, pp. 253–265. Springer International Publishing, 2015.

Janzing, Dominik, Balduzzi, David, Grosse-Wentrup, Moritz, Schölkopf, Bernhard, et al. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013.

Kandasamy, Kirthevasan, Krishnamurthy, Akshay, Poczos, Barnabas, and Wasserman, Larry. Nonparametric von mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems*, pp. 397–405, 2015.

Khan, Shiraj, Bandyopadhyay, Sharba, Ganguly, Auroop R, Saigal, Sunil, Erickson III, David J, Protopopescu, Vladimir, and Ostrouchov, George. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*, 76(2):026209, 2007.

Kozachenko, LF and Leonenko, Nikolai N. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

Krishnaswamy, Smita, Spitzer, Matthew H, Mingueneau, Michael, Bendall, Sean C, Litvin, Oren, Stone, Erica, Pe'er, Dana, and Nolan, Garry P. Conditional density-based analysis of t cell signaling in single-cell data. *Science*, 346(6213):1250689, 2014.

Loader, Clive R et al. Local likelihood density estimation. *The Annals of Statistics*, 24(4):1602–1618, 1996.

Lombardi, Damiano and Pant, Sanjay. Nonparametric k-nearest-neighbor entropy estimator. *Physical Review E*, 93(1):013310, 2016.

Moddemeijer, Rudy. On estimation of entropy and mutual information of continuous distributions. *Signal processing*, 16(3):233–248, 1989.

Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 2015.

Mueller, Jonas, Jaakkola, Tommi, and Gifford, David. Modeling trends in distributions. *arXiv preprint arXiv:1511.04486*, 2015.

Pál, Dávid, Póczos, Barnabás, and Szepesvári, Csaba. Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Advances in Neural Information Processing Systems*, pp. 1849–1857, 2010.

Paninski, Liam. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

Pearl, Judea. *Causality*. Cambridge university press, 2009.

Póczos, Barnabás, Xiong, Liang, and Schneider, Jeff. Nonparametric divergence estimation with applications to machine learning on distributions. *arXiv preprint arXiv:1202.3758*, 2012.

Rényi, Alfréd. On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451, 1959.

Richardson, Robin J and Evans, Thomas S. Nonparametric causal models. 2015.

Shajarisales, N., Janzing, D., Schölkopf, B., and Besserve, M. Telling cause from effect in deterministic linear dynamical systems. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 285–294. JMLR, 2015.

Shannon, C.E. A mathematical theory of communication. *Bell System Tech. J.*, 27:379423 and 623656, 1948.

Sheather, Simon J and Jones, Michael C. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 683–690, 1991.

Sricharan, Kumar, Raich, Raviv, and Hero III, Alfred O. Empirical estimation of entropy functionals with confidence. *arXiv preprint arXiv:1012.4188*, 2010.

Tsybakov, Alexandre B and Van der Meulen, EC. Root-n consistent estimators of entropy for densities with unbounded support. *Scandinavian Journal of Statistics*, pp. 75–83, 1996.

Wang, Qing, Kulkarni, Sanjeev R, and Verdú, Sergio. Divergence estimation for multidimensional densities via-nearest-neighbor distances. *Information Theory, IEEE Transactions on*, 55(5):2392–2405, 2009.