On the Theory and Algorithm for rigorous discretization in applications of Information Theory

Venkateshan Kannan, Jesper Tegnér *

Computational Medicine Unit, Karolinska Institute, Stockholm, Sweden 17176

Abstract

We identify fundamental issues with discretization when estimating information-theoretic quantities in the analysis of data. These difficulties are theoretical in nature and arise with discrete datasets carrying significant implications for the corresponding claims and results. Here we describe the origins of the methodological problems, and provide a clear illustration of their impact with the example of biological network reconstruction. We propose an algorithm (shared information metric) that corrects for the biases and the resulting improved performance of the algorithm demonstrates the need to take due consideration of this issue in different contexts.

This is investigated in the context of network inference [1–5] where information-theoretic methods have been extensively applied [6–9]. Despite its ubiquity, general systematic analysis of the techniques and their underlying assumptions have been few [10].Although information entropy, and its variations such as the joint or conditional entropy, is well-defined for discrete variables, its formulation for continuous probability distribution is far from unambiguous. The flexibility in this definition lies at the heart of the issue when applying it consistently to different physical systems. Specifically, entropy estimation in continuous cases requires discretization of the variable values, and the calculation results are not invariant to the choice of discretization. With the focus on reverse-engineering, we describe the fundamental issues with information-theoretic methods, construct examples that highlight them, and propose an algorithm to remedy the situation.

1

1 Results

Mathematically, Shannon entropy of a variable X, defined as $H(X) = \sum_i -p_i \log(p_i)$ [11], is a well-defined quantity when X is discrete, taking distinct values i with probabilities p_i. This is not the case when X is sampled from a continuous probability distribution where an equivalent definition would just give infinity [12]. Nonetheless, there exists a generalization of mutual information (MI) of two variables X and Y with joint probability distribution $P_{X,Y}(u,v)$ and corresponding marginals $P_X(u)$ and $P_Y(v)$:

$$\begin{aligned} I &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{X=u,Y=v} P_{X,Y}(u,v) \log \frac{P_{X,Y}(u,v)}{P_X(u)P_Y(v)} \end{aligned} \tag{1}$$

from discrete to continuous joint probability distribution

$$I = \int du\,dv\, P_{X,Y}(u,v) \log \frac{P_{X,Y}(u,v)}{P_X(u)P_Y(v)} \tag{2}$$

which is well-behaved [12]. However, given discrete ordered pair (x_p, y_p) sampled from an unknown joint probability distribution over continuous variables, we would have to approximate the true (continuous) distribution from it. At the level of probability, this is achieved by binning and discretizing the range of values of the variables.

While this is carried out straightforwardly, there are pitfalls in simple generalization to mutual information. The key problem is that mutual information depends in a significant way on the discretization parameters, even when the number of points is very large. We show this with the following simple but general example. Consider ordered pairs (x_v, y_v), $v = 1, 2, \cdots N$ from a joint distribution with the only requirement that $x_u \neq x_v$ for $u \neq v$ and likewise for y's. If the entire range of X and Y is considered as a single segment for discretization, $\Delta_X^{(0)}$ and $\Delta_Y^{(0)}$ then the joint probability distribution $P_{X,Y}(\Delta_X^{(0)}, \Delta_Y^{(0)}) = 1$ and hence the marginals are also unity for the corresponding X and Y ranges. It can be immediately seen that the mutual information is zero. Now, for the other extreme, we choose the interval widths δ_X and δ_Y such that there is at most one point lying within each interval (for both the variables). We can always do this by selecting $\delta_X < \min_{u \neq v, u,v=1,2,\cdots,N}\{|x_u - x_v|\}$ and $\delta_Y < \min_{u \neq v, u,v=1,2,\cdots,N}\{|y_u - y_v|\}$. In that case, the discretized joint probability in each rectangular cell is $1/N$ if it is occupied and 0 otherwise. In much the same way the marginal probability in each interval is either $1/N$ or 0 for both X and Y. Fig (1) shows how would be done in the general case (note that interval widths there are not uniform). We have then the mutual information:

$$I = \sum_{\alpha,\beta} P_{X,Y}(\alpha,\beta) \log \frac{P_{X,Y}(\alpha,\beta)}{P(\alpha)P(\beta)} = \sum_{\alpha,\beta} 1/N \log \frac{1/N}{(1/N)(1/N)} = \log N \tag{3}$$
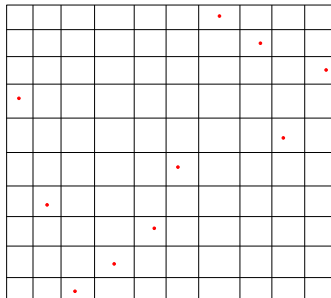
2

Figure 1: *Grids in 2D space for a set of 10 points such that for any partition along the X-direction (Y-direction) there is exactly one point within it. This corresponds to a maximum possible mutual information of* $\log 10$.

Thus, for a set of N observations from a joint probability distribution, the mutual information can vary between 0 and $\log N$ depending on the size of the binning. What is most striking about this is that *this is true regardless of the true mutual information of the underlying joint probability distribution.* As the upper bound grows with the size N, having a larger sample number does not solve the fundamental problem of the inherent ambiguity of defining mutual information using discrete data points.

1.1   Standard partitioning biases

Recognizing this basic limitation, we wish to consider the implications of different choices of binning in estimating mutual information [13]. We want to understand the underlying biases and the possible deviation of mathematical properties from that of the original definition.

In general, mutual information increases as we increase the number of partitions of the space. We can see that in Fig (2) the variation of the estimated mutual information for two multivariate Gaussian random variables with the number of bins (assuming fixed bin length) for different sample sizes N. The plotted values for the estimated MI comes from the average taken over 50 samples. In the supplement, we prove a general result stating that the doubling of partition size would increase the mutual information between two variables for an arbitrary set of samples. However, this by itself would not be an issue if we are considering reverse engineering in networks as long as the rankings of estimated mutual information between variables is preserved with respect to their true values

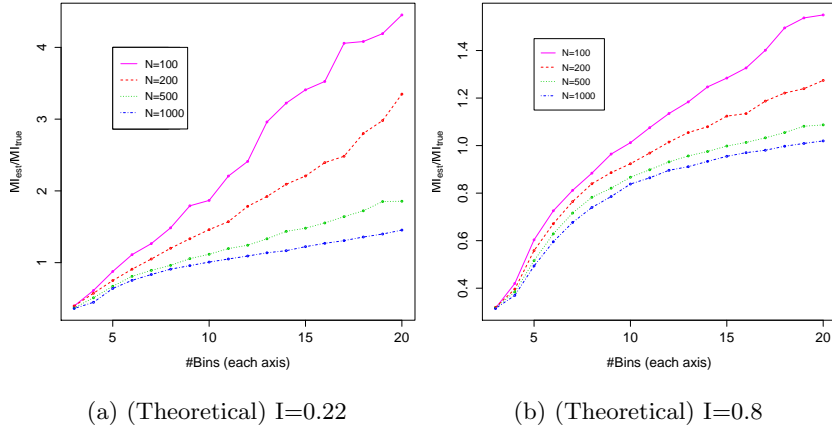(a) (Theoretical) I=0.22   (b) (Theoretical) I=0.8

Figure 2: *Variation of the estimated entropy as a function of number of bins (along each axis) for a pair of jointly Gaussian random variables.*

There are other biases and errors that are introduced by the choice of the partitioning scheme that we will consider here. Broadly, partitioning of the two dimensional space of variables X and Y fall into two general categories: uniform number of bins (b) where the number of divisions of both axis is identical, uniform width where the width of each partition (w) is identical for both variables, and uniform frequency where the number of points falling into each partition (for each variable) variable is fixed.. We will now turn to each of these specific methods.

1.1.1 Uniform number of partitions

Here each axis of the space is split into equal number of uniform segments. The advantage of this method is that it works well when the data is distributed reasonably evenly across the range of each axis which happens when the density of points falls off sharply outside of the confined region under consideration.

However, the difficulty arises with probability densities that have a fat tail, where the spread is wide but the majority of points are still lying within a smaller region. A binning of this sort would then be finer for points in the periphery but coarser in the denser regions, thus tending to underestimate the mutual information from the latter areas.

We demonstrate this by estimating the entropy for two samples , the first of which comes from the exponential family,

$$P_{X,Y} = \frac{1}{\Gamma(\theta)\lambda} x^\theta e^{-\frac{xy}{\lambda}-x} \tag{4}$$

with $\theta = 4, \lambda = 5$ and the second a multivariate Gaussian in two-variables

$\sim e^{-(X-\mu_x)\cdot\Sigma^{-1}\cdot(Y-\mu_y)}$ where

$$\Sigma = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$$

and $(\mu_x, \mu_y) = (5, 5)$. Each set consists of 200 points, and the barplots of distributions (Fig.(3a,3b)) across a uniform 10×10 grid show that while the multivariate normal is spread quite smoothly across the entire surface, it falls off in the other for $X, Y > 5$.

The mutual information can be theoretically computed in these two cases:

$$I^{exp}(\theta, \lambda) = \lambda(\psi(\theta + 1) - \log\theta)$$

where ψ is the digamma function and

$$I^{Gauss}(\Sigma) = \frac{1}{2}\log\frac{\sigma_{xx}^2\sigma_{yy}^2}{\sigma_{xx}^2\sigma_{yy}^2 - \sigma_{xy}^2}$$

which evaluate to $I^{exp} = 0.79$ and $I^{Gauss} = 0.51$ but the estimated mutual information from the above binning leads to $I_e^{exp} = 0.21$ and $I_e^{Gauss} = 0.59$.

The reason for the severe underestimation in the first case is precisely the fact that there is a closer clustering of points within a smaller region where this type of binning tends to be 'too wide' and averaging the finer distinctions. In the multivariate normal case, this difference is not so striking, and the estimation is far more accurate.

1.1.2   Uniform partition widths

The other standard approach to choosing a grid on the two dimensional space of points is to set the width of every partition on both axes to be a constant. As the spread of the two variables is in general different, the total number of partitions need not be the same. The idea behind this is that the width represents the distance over which points are considered to fall within the same discrete category. It may be set based on estimates of noise in the sample, or, where available, prior knowledge of the form of the joint distribution.

There is however a fundamental problem with this approach : setting the widths to be a constant destroys the scaling invariance of the theoretical definition of mutual information for continuous variables. In Eq. (2), the integral is invariant under the transformation $y \to cy$, $P_{X,Y}(x, y) \to cP_{X,Y}(x, cy)$, $P_Y(y) \to cP_Y(cy)$ which corresponds to scaling the metric of the Y-axis by a constant. It must be noted that this is not an incidental property of the integral but one that is central to the usefulness of mutual information, i.e., mere scaling of the underlying space should not, and does not, alter the relative information between two variables.
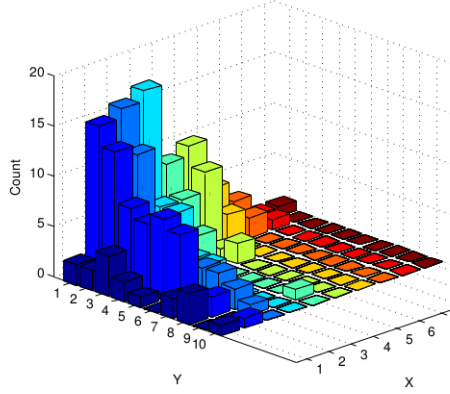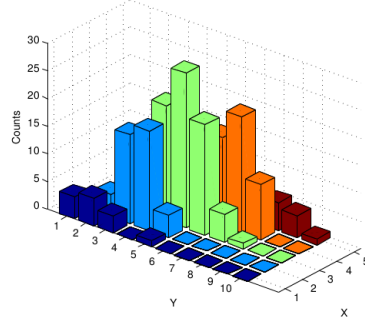
(a) *Sample from an exponential family Eq. (4) with $\theta = 3$ and $\lambda = 5$. Theoretical (estimated) mutual information is 0.79 (0.21)*
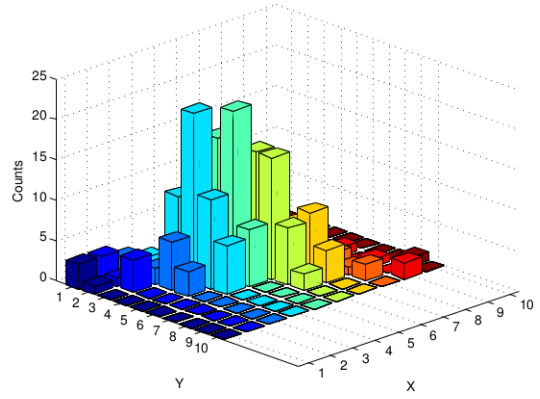
(b) *Normal distribution with $\sigma_{xx}^2 = \sigma_{yy}^2 = 5, \sigma_{xy} = 4$. Theoretical (estimated) mutual information is 0.51 (0.59)*
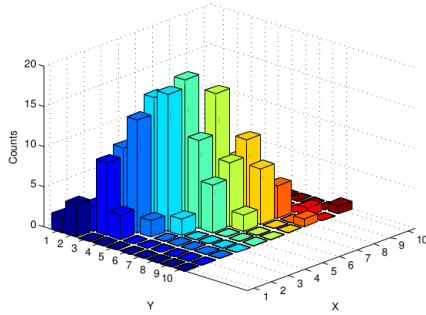
(c) *Normal distribution with $\sigma_{xx}^2 = 5, \sigma_{yy}^2 = 5, \sigma_{xy} = 4.5$. Theoretical (estimated) mutual information is 0.83 (0.79)*

(d) *Normal distribution with $\sigma_{xx}^2 = 5, \sigma_{yy}^2 = 2.5, \sigma_{xy} = 3.18$. Theoretical (estimated) mutual information is 0.83 (0.67)*

Figure 3: **Mutual Information and meshing: In (a) and (b), the grid choice turns the space of points into a square with equal number of partitions along both, while for (c) and (d) the width of every partition is held fixed.**
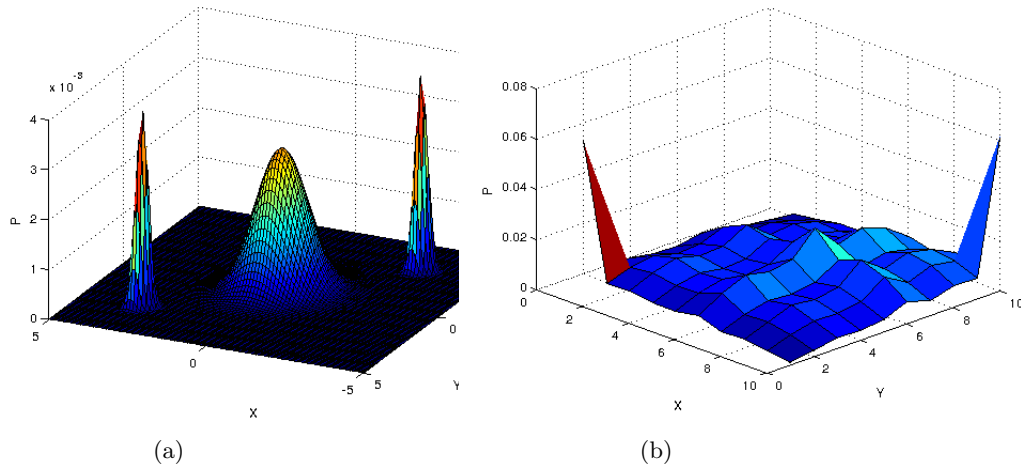
6

Figure 4: (Left) *Mixed-Gaussian distribution.* (Right)*Equal-frequency partitioning skews the distribution. Estimated(Numerical) mutual information is 0.19 (0.45)*

Choosing binning styles with equal widths implies that scaling the points along one direction in the space would change the number of partitions on its axis, which changes the estimated mutual information. The effect of this can be easily observed by considering samplings from two multivariate normal distributions in two dimensions that differ only by the scaling of the y-axis by a factor of $1/\sqrt{2}$(Fig. (3c) and Fig. (3d). By definition, the scale-invariance implies equal mutual information but the requirement of setting equal widths (unity in this case) leads fewer bins in the second case and consequent underestimation of its value.

1.1.3   Equal Frequency Partitioning

In this case, the boundaries of the partitions for a given variable are chosen such that each division contains an equal number of points. The general advantage of this method is that, unlike the above two, every row or column in the grid would contain equal number of points, eliminating the problem highlighted in Fig (3a,3b). However, this would skew joint distributions that have strong association between two variables to a more uniform shape, causing an underestimation in mutual information. We can see this with the mixed-Gaussian distribution with three modes (Fig. (4)) where the effect of choosing the equal frequency partitioning on a 10×10 grid flattens the peak in the center spreading the probability over a wider region, leading to a decrease in mutual information.

It is thus clear from the examples above that the standard types of partitioning is in general unsatisfactory in terms of obtaining reasonable and consistent mutual information scores. It should be noted that although our examples consider

7

discrete binning of the data, the kernel-based estimation of mutual information [14, 15] has the same fundamental problem. While it is not as pronounced as it is with direct binning, the width of the kernel in that case is equivalent to the width of the bins here (equal-width partition). The same is true with estimators like the k-nearest neighbors [16] where the scores depend on the parameter corresponding to the number of neighbors.

1.2   Adaptive Partitioning for Network Inference

Data-driven reverse-engineering of genetic networks uses pairwise association between genes to determine true interactions among them. Information theory is applied effectively for this task by estimating the mutual information between every pair of network node variables and reconstructing the network based on the relative rankings of the different edge scores [8, 9, 17]. A distinct advantage being that mutual information detects all forms of associations while correlation-based measures perform well only with linear relations.

We propose a novel method of partitioning of space in the context of network reverse-engineering that aims to reduce the biases created by the standard techniques. This includes choosing grid sizes that take into consideration (a) overall spread of the data among all variables (b) spread of the values of the pair of node variables in question (c) dependence of numerical estimation of mutual information on this spread (d) appropriate normalization using the entropy of each variable.

Given set V of N node variables and M samples, we choose first a standard partition width $w_0 = w_{int}/K_{min}$, where $w_{int} = \text{median}\{\sigma(X)|X \in V\}$, and K_{min} is an integer parameter that represents the minimum number of partitions and $\sigma(X)$ is the empirical standard deviation of the X variable values.

For any two node variables X and Y whose mutual information we want to estimate, the number of bins b_X and b_Y are obtained using the following algorithm (see Fig. 5). In addition to K_{min}, we have another integer parameter K_{max} that sets the maximum number of partitions.

Steps 1-3 may look complicated but the idea is simple enough: we prefer to maintain the bin widths constant as long as their numbers lie between K_{min} and K_{max}. This is primarily in recognition of the fact that spread of the node variable is a measure of the strength of the interactions: if the values are nearly the same, the uniform width approach would select fewer partitions. However, to ensure that the distribution is not skewed when the spread of one or both is large (or too small), we need to correspondingly *rescale both variables* to force them to lie within that range. Heuristically, the two limits K_{min} and K_{max}

Step 1: b_X is first set to $[\frac{\sigma(X)}{w_0}]$ where $[.]$ refers to the nearest integer function. Likewise for b_Y. Assume $b_X \leq b_Y$.

Step 2 : If $b_Y > K_{max}$, then reset $b_X = [b_X(K/b_Y)]$ and $b_Y = K$. If the new $b_X < K_{min}$, reset $b_X = K_{min}$.

Step 3: If $b_X < K_{min}$, then $b_Y = [\max\{K_{min}, \min\{K_{max}, K_{min}\sqrt{b_Y/b_X}\}\}]$ and $b_X = K_{min}$.

Step 4: Once the binning is fixed, we proceed to calculate the number of points falling within each rectangle, and the discretized form of mutual information between the two variables.

Step 5: We normalize the mutual information by dividing by $\min\{H(X), H(Y)\}$, where $H(X)$ and $H(Y)$ are the entropies of X and Y calculated using the same bin numbers b_X and b_Y respectively. We call the resulting quantity shared information metric (SIM).

Figure 5: **Adaptive Paritioning Algorithm**

ensure that the relative values of mutual information are not under- or over-estimated.

Normalization Measure:
Step 5 was motivated by two considerations. First, we note that a proper normalization is required when having different bin sizes. If, for example, there is a node variable with a small spread but happens to have a strong direct interaction with another node, the binning scheme would then 'overcorrect' for the small spread. We overcome that by such a normalization scheme. Second, we recognize that what is significant in reverse-engineering is not the absolute value of mutual-information but the MI relative to the information content of the two node variables. The inequality $I(X,Y) \leq \min\{H(X), H(Y)\}$ captures the fact that this quantity represents that part of the information that is shared between the nodes. The fraction of what is shared should serve as a better indicator of true interaction than their absolute values.

1.3 Evaluation on networks

We assess the algorithm by studying its performance on *in-silico* networks against that of standard methods (see Section 2 for more details). The initial comparison is with respect to the estimation method using only uniform bin size or uniform bin number. Evaluation of the performance is done by plotting the precision-recall curve (Fig. (6)) (see Section 2 for more details on their definition). It can be clearly seen that, at any given value of recall, the precision is significantly higher when using adaptive non-uniform discretization. We also consistently found better results regardless of the size or the topology of the network or its dynamics.

Our method performed better even against standard mutual-information based approaches such as kernel estimation and DPI (ARACNE) [15] and the k-nearest neighbor algorithm (with DPI) [16] (Fig. (7)).
For biological data sets, we considered the gene regulatory network of *E. coli*, and the expression data for it from the DREAM 5 challenge [18]. In this case, the set of transcription factors is known, and the aim is to determine their targets. Thus our binning choice should reflect this prior knowledge and the algorithm was modified accordingly (described in the supplement). Of the set of 2000 interactions we evaluated, of which only 111 represented true regulatory interactions, our method clearly performed better than the standard method of discretization (Fig. 8).

2 Methods

We constructed synthetic networks obeying power-law degree-distribution with the R package *igraph*. The data was generated by carrying out numerical integration of the Hill-kinetic equations corresponding to the network topology until global steady state was reached. Each sample in the data set corresponds to the steady state value of all variables. Different samples correspond to steady states of different equations obtained with set of parameters comes from a uniform distribution.

The precision-recall curve (PRC) is used to evaluate these networks, where precision and recall are given by:
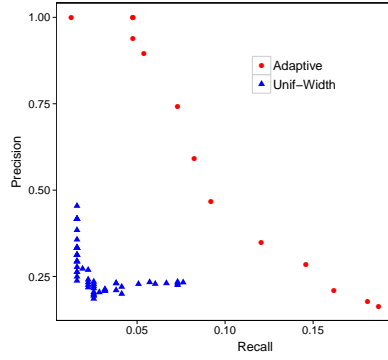
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

and TP= True Positives (the number of correct interactions identified), FP= number of pairs incorrectly identified, and FN=number of interactions that were
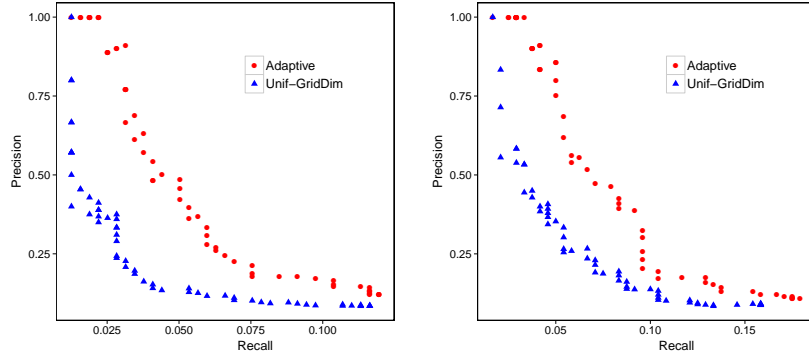
(a) $K_{min} = 4$ and $K_{max} = 10$. $N = 100$ and $E= 160$

(b) K_{min}=2 amd K_{max} =5. $N = 100$ and $E= 120$

(c) Equal-width vs Adaptive Discretization: $N = 100$ and $E = 160$

Figure 6: *Comparison of the performance of network edge determination using uniform number off bins (10 and 5 in figure (a) and (b) respectively) or uniform bin width (Fig. (c)) against a non-uniform discretization. N and E refer to the total number of nodes and true edges in the network.*
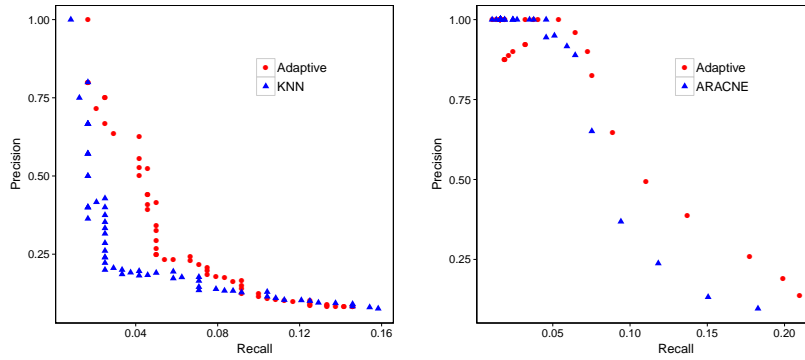
(a) $N = 100$ and $E = 120$ (b) $N = 100$ and $E = 185$

Figure 7: *Comparison of the performance of network edge determination using k-nearest neighbor method (with DPI) (Fig. (a)) and ARACNE (Fig (b)) against the non-uniform discretization with the shared information mertic (SIM). N and E refer to the total number of nodes and true edges in the network.*
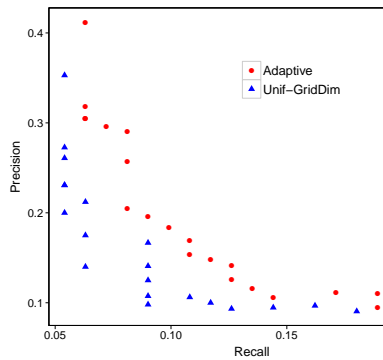

Figure 8: *Comparison of the performance of identification of targets of transcription factors of the gene regulatory network of E. coli. Expression data and true interaction set obtained from the DREAM 5 challenge.*

unidentified. In general, there is a trade-off between precision and recall, as we slide the threshold parameter.

3 Conclusions

We have described and detailed significant problems with the definition and estimation of mutual information, a quantity that is central to information theory. We argue that this issue is of a theoretical nature leading to an inherent arbitrariness in its estimation from a discrete set of points. Moreover, the the errors introduced by the standard forms of partitions of space is explicitly demonstrated with examples .

By formulating a novel adaptive partitioning algorithm for reverse engineering of networks, we further show the impact of these estimation biases in a real-world context where mutual information is applied extensively. The superior performance of the method over standard approaches on both in-silico and real biological networks is not only an advancement in that field, but also clearly implies the necessity to carefully understand the problems with estimating mutual information.

The applications of information theory to various fields have grown enormously in recent decades as the concepts have become central to areas such as physics, communications and signaling, inference theory, multiple biological sciences, pattern recognition and artificial intelligence [19–24]. We believe that a thorough investigation of the fundamental issues involved in the subject would be immensely beneficial to all of their applications.

Appendix A   Proof of increase in mutual information

We show that doubling the partition number along one of the directions leads to a non-negative change in the mutual information. Starting with a division of the 2D space into $L \times M$ blocks we bisect each of the LM units along the $X-$axis to create $2L \times M$ cells. The mutual information:

$$I = \sum_{i=1}^{L} \sum_{j=1}^{M} P_{X,Y}(i,j) \log \frac{P_{X,Y}(i,j)}{P_X(i)P_Y(j)} \tag{5}$$

where $P_{X,Y}(i,j)$ is the fraction of particles lying in the rectangular cell (i,j). After bisection, we have a new probability distribution $P^{2L}(i,j)$ (dropping the double-index subscript) where $i = 1, 2, \cdots 2L$, such that $P^{2L}(2k-1,j) + P^{2L}(2k,j) = P(k,j)$ for $k = 1, 2, \cdots L$. We will show that for every original cell, its contribution to MI is exceeded by the sum of the contributions of the

bisected components and thus obtain a stronger version of our required result. To prove:

$$P^{2L}(2k-1,j)\log\frac{P^{2L}(2k-1,j)}{P_X^{2L}(2k-1)P_Y(j)}+P^{2L}(2k,j)\log\frac{P^{2L}(2k,j)}{P_X^{2L}(2k)P_Y(j)} \geq P(k,j)\log\frac{P(k,j)}{P_X(k)P_Y(j)}$$
(6)

which reduces to,

$$P^{2L}(2k-1,j)\log\frac{P^{2L}(2k-1,j)}{P_X^{2L}(2k-1)}+P^{2L}(2k,j)\log\frac{P^{2L}(2k,j)}{P_X^{2L}(2k)} \geq P(k,j)\log\frac{P(k,j)}{P_X(k)}.$$
(7)

Substituting $p_a = P^{2L}(2k-1,j), p_b = P^{2L}(2k,j), p_1 = P_X^{2L}(2k-1), p_2 = P_X^{2L}(2k)$,

$$p_a\log\frac{p_a}{p_1} + p_b\log\frac{p_b}{p_2} \geq (p_a+p_b)\log\frac{p_a+p_b}{p_1+p_2}$$
(8)

Straightforward rearrangements of the terms leads to,

$$\frac{p_a}{p_a+p_b}\log\frac{p_a}{p_a+p_b} + \frac{p_b}{p_a+p_b}\log\frac{p_b}{p_a+p_b} - \frac{p_a}{p_a+p_b}\log\frac{p_1}{p_1+p_2}$$
$$- \frac{p_b}{p_a+p_b}\log\frac{p_2}{p_1+p_2} \geq 0$$
(9)

which is equivalent to demonstrating that :

$$x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y} \geq 0$$
(10)

for $0 \leq x,y \leq 1$. The left hand side is nothing but KL-divergence between two distributions on two element set with probabilities $\{x, 1-x\}$ and $\{y, 1-y\}$, which is always positive.

Appendix B   Modified Algorithm for determining targets

We used a modification of the general algorithm for the TF-target identification in the *E. coli* expression data set. As the set of transcription factors are known, the binning adjustments can be limited to the targets here and thus our technique is significantly simplified. As before $w_0 = w_{int}/H$, where $w_{int} = \text{median}\{\sigma(X)|X \in V\}$, and H and K represent the minimum and maximum number of bins.

Step 1: Fix $b_X = (H+K)/2$
Step 2 : If $b_Y > K$, then reset $b_Y = K$.
Step 3: If $b_Y < H$, then $b_Y = H$.
Step 4: Once the binning is fixed, we proceed to calculate the number of points falling within each rectangle, and the discretized form of mutual information between the two variables.

Step 5: We normalize the mutual information by dividing by $\min\{H(X), H(Y)\}$, where $H(X)$ and $H(Y)$ are the entropies of X and Y calculated using the same bin numbers b_X and b_Y respectively.

References

[1] Yeung M.K., Tegnér J., Collins J.J., **Reverse engineering gene networks using singular value decomposition and robust regression**, *Proc Natl Acad Sci USA*, **99(9)** :6163-6168 (2002).

[2] Hecker M.,Lambeck S., Toepfer S.,van Someren E., Guthke R., **Gene regulatory network inference: Data integration in dynamic modelsA review**, BioSystems **96** 86103 (2009)

[3] Markowetz F.,Span R., **Inferring cellular networks a review**, *BMC Bioinformatics* **8** (Suppl 6):S5 (2007)

[4] Bansal M.,Belcastro V., Alberto Ambesi-Impiombato1, di Bernardo D., **How to infer gene networks from expression profiles**, *Molecular Systems Biology* **3**:78 (2007)

[5] Hendricks D.M., Hendriks M.M. W. B.,Eilers P. H. C.,Smildeac A.K., Hoefslootac Huub C. J.,**Reverse engineering of metabolic networks, a critical assessment**, *Mol. BioSyst.*, **7**, 511 (2011)

[6] Liang S, Fuhrman S and Somogyi R., **Reveal, a general reverse engineering algorithm for inference of genetic network architectures**, *Pac Symp Biocomput*, 18 (1998)

[7] Butte A.J.,Kohane I.S., **Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.**, *Pac Symp Biocomput.*,**5** 415 (2000)

[8] 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* **334** (6062): 1518 (2011)

[9] Faith J.J., Hayete B.,Thaden J.T.,Mogno I., Wierzbowski J., Cottarel G.,Kasif S., Collins J.J., Gardner T.S., **Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles**, *PLoS Biol* **5(1)**: e8.

[10] Margolin A.A., Wang K., Califano Nemenman I, **Multivariate dependence and genetic networks inference**, *IET Syst Biol.* **4(6)** :428

[11] Shannon C.E., **The mathematical theory of communication** ,*Bell Syst. Techn. Journal*, **27**, 379 (1948).

[12] Lesne A.,**Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics**, *Mathematical Structures in Computer Science*, **24**, SI 03, 24031 (2014)

[13] de Matos Simoes R., Emmert-Streib F., **Influence of Statistical Estimators of Mutual Information and Data Heterogeneity on the Inference of Gene Regulatory Networks** *PLoS ONE* **6(12)**: e29279 (2011)

[14] Moon Y.I., Rajagopalan B., Lall U.,**Estimation of mutual information using kernel density estimators**, *Phys. Rev. E* **52**, 2318 (1995)

[15] Margolin A.A., Nemenman I., Basso K., Wiggins C., Stolovitzky G., Dalla Favera R., Califano A., **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context**, *BMC Bioinformatics* **7** (Suppl 1):S7

[16] Kraskov A.,Stgbauer H., Grassberger P., **Estimating Mutual Information**, *Phys. Rev. E* **69**, 066138 (2004)

[17] Chow CK, Liu CN, **Approximating discrete probability distributions with dependence trees**,*IEEE Trans. Inf. Thy.*,**14(3)**, 462 (1968)

[18] Marbach *et al*, **Wisdom of crowds for robust gene network inference**, *Nature Methods*, **9**, 796 (2012)

[19] Jaynes E.T., **Information Theory and Statistical Mechanics**, *Phys. Rev.*, **106**, 620 (1957)

[20] Adami C., **Information theory in molecular biology**, *Physics of Life Reviews*, **1(1)**, 3 (2004)

[21] Maasoumi E., **A compendium to information theory in economics and econometrics**, **12 (2)**, 137 (1993)

[22] Pereira F., **Formal grammar and information theory: together again?** , *Phil. Trans. R. Soc. Lond. A* , **358**, 1239 (2000)

[23] Tishby N., Pereira F.C., Bialek W., **The Information Bottleneck method**, *The 37th annual Allerton Conference on Communication, Control, and Computing*, 368377 (1999)

[24] Studholme C., Hill D.L.G., Hawkes D.J., **An overlap invariant entropy measure of 3D medical image alignment** *Pattern Recognition* **32(1)** 7186 (1999)