 doi:10.20944/preprints202202.0254.v1

# RESOLUTION LIMIT IN STATISTICAL INDEPENDENCE AND BAYESIAN NETWORK SCORING FUNCTIONS

GRIGORIY GOGOSHIN [1] AND ANDREI S. RODIN [2]

Abstract. In this paper we consider the congruence problem that arises in the post-analysis of Bayesian network models reconstructed from different datasets. Apart from the structure, a typical network numerically encodes relationship intensities, assigning numerical score to network edges via the scoring criterion used in the reconstruction process. This scoring is rarely a directly interpretable quantity with proper units of measure and an absolute scale, and often comes short in desirable characteristics of a true metric. This leads to poor portability of edge magnitude considerations between similar networks, originating from different sources. In this work, we address this problem by estimating the effect that data-specific resolution limit has on conditional independence, as reflected by information-theoretic entropy, and by the appropriate modification of MDL score, which removes the inconsistency between the score components in both the meaning and units. We also numerically validate our findings and expose additional performance advantages obtained via this modification.

## Introduction

Probabilistic Bayesian network (BN) modeling is an indispensable tool in modern medical and life science. Apart from the standard array of data-analytic uses that are common to all probabilistic models, it has the advantage of capturing the complex structure of the web of relationships underlying the biological reality. When used for extraction of meaning from data, BN reconstruction generates valid, evidence-based, directly interpretable hypothesis, which does so much more for both the practical and the theoretical research than numerical replication of phenomenology alone can hope to do, however accurate it may be.

BN-based dependency modeling has established itself firmly in computational biology and gained significant traction in biomedical data analysis. It finds applications in a wide variety of areas ranging

[1,2] Department of Computational and Quantitative Medicine, Beckman Research Institute, and Diabetes and Metabolism Research Institute, City of Hope National Medical Center, 1500 East Duarte Road, Duarte, CA 91010 USA

*E-mail addresses*: [1]ggogoshin@coh.org, [2]arodin@coh.org.

*Key words and phrases.* Bayesian networks, probabilistic networks, conditional independence, MDL, BIC, information-theoretic entropy.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2022 doi:10.20944/preprints202202.0254.v1

from flow cytometry [1] to chromatin interactions [2], from molecular evolution [3] to genomics [4, 5, 6, 7], transcriptomics [8, 9] and even metabolomics [10].

But BN reconstruction itself is not without its own set of difficulties. We will not descend into the bottomless pit of complications that have to be overcome in order to obtain an adequate for practical use result, suffice it to say that the computational difficulties are significant. Here, we would like to concentrate on issues that meet the eye at the "user interface" level of BN modeling.

The specific problem that we want to address arises in the context of analysing BN networks coming from different sources. Even when any two BN networks span the identical set of variables, comparing them is a non-trivial matter due to the fact that certain structural features may belong to classes of equivalence. But, the task becomes even more complicated when it comes to the assessment of variable dependence intensities, and the role they play in the rest of the network. The situation is exacerbated by the fact that the edge intensities are typically evaluated by scoring criteria, used as a part of the structure recovery process, that offer no more than relative ordering of edges on some arbitrary scale, rendering the interpretation of these relative intensities difficult and non-portable.

The outlined issue motivates designing a scoring criterion that could serve not only as an objective function for structure recovery, but as measure of intensity or strength on an absolute scale, possibly with some desirable properties of a distance function/metric, allowing direct comparison of edges, edge bundles and paths between networks, without digging through massive conditional probability tables and factorizations in search for explanations for local network behavior. We will proceed by first considering possible modifications of a well-known and reliable Minimum Description Length (MDL) criterion that also appears to be equivalent to another well-known Bayesian Information Criterion (BIC) [11, 12]. For a graph $G$

$$MDL(G) = LL(G) - \frac{1}{2}C(G)\log(N)$$

where the term

$$LL(G) = N\sum_i H(X_i|\pi_i)$$

is the log-likelihood of $G$ written in terms conditional entropy $H$ of individual nodes $X_i$ and their parent node sets $\pi_i$, and the term $\frac{1}{2}C(G)\log(N)$ represents the description length of $G$, where $C(G)$, called complexity, is usually taken to be proportional to the number of free parameters necessary to represent the factorization of the joint probability of $G$. While MDL performs well as an objective function, its composition gets somewhat in the way of interpreting the scores that it could generate for our purpose.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2022 doi:10.20944/preprints202202.0254.v1

First of all, log-likelihood term is proportional to the sample size $N$, which means that for every new dataset this portion of the score is bound to a different scale unless the sample size remains constant. Second, the description length term is itself on a different scale than the log-likelihood, meaning that the relative contribution to the total score varies at a different rate between the two terms across different datasets. To summarize with a specific use case example, given some data, a network obtained via subsampling would be numerically not comparable to the network obtained using the whole data, crippling certain aspects of robustness and stability analysis.

To overcome this obstacle, the most obvious line of thinking is to do away with the sample size dependence. Conveniently, $LL(G)$ can be easily rescaled to the sum of local conditional entropies. But simply dividing $N$ out results in the appearance of $1/N$ in the description length term, and while the first term is, at least, interpretable, the second term is still incongruent in the sense that it not measured in the same units as entropy, isn't proportional to it, and seems to be on a different scale altogether, so things aren't so simple. Further, why are these seemingly incongruent terms appear together? The measure of complexity, which is what the description length essentially is, is neither proportional to the measure of entropy, nor does it stand in a one-to-one correspondence to it [13].

Eventually, it becomes apparent that the description length acts as an ad-hoc penalty term chosen to control complexity of the network that arises due to the tendency of log-likelihood or, equivalently, conditional entropy to overfit when the search algorithm seeks an optimal structure. This overfitting, however, is an indication of misalignment between the objective of the optimization procedure, relying on log-likelihood or entropy, with what it should be trying to achieve. To be more precise, minimizing the conditional entropy or the log-likelihood , which is what MDL essentially engages in, isn't quite the core objective, even if it partially coincides with the goal of finding an optimal structure. This is apparent from the fact that the conditional entropy minimum is achieved in the situation where every sample gets its own class, or where the joint events of the ancestor variables are fine enough to be close to homogeneous, which clearly doesn't have to coincide with the true solution. Even more importantly, the true solution should correspond to the correct conditional probability distribution, as opposed to the distribution of minimum entropy. On the other hand, perhaps the most important detail about MDL is that the conditional entropy minimization serves the purpose of finding locally dependent variables via an update of the form

$$\Delta H = H(X|\pi, Y) - H(X|\pi)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2022 doi:10.20944/preprints202202.0254.v1

which is essentially an independence test, in the sense that $\Delta H = 0$ when $X$ is independent of $Y$ given its ancestor set $\pi$. This gives us the hint that we don't have to worry so much about the justification of entropy application and its correspondence with the true solution, since we may just be able to get away with always maximizing the local dependence without running into an over-fitting problem by maintaining a stringent independence cutoff. With all of the above in mind, we are ready to rederive the score starting from these basic principles in such a way, so as to make all the terms contextually congruent to each other and independent of sample size. For now we will accept the general form of the objective function to be

$$S(G) = \sum_i (H(X_i|\pi_i) - \xi(X_i))$$

so that the local iteration update is

$$\Delta S(X_i) = H(X_i|\pi_i, Y) - H(X_i|\pi_i) - \Delta\xi(X_i)$$

where the second term $\Delta\xi$ should perform the function of independence filter, and $\xi$ should reflect local independence policy. This will be made clearer in the next section.

## Methods: Numerical satisfiability and resolution limit in independence criteria.

One of the difficulties in assessing independence lies in the fact that analytical criteria, such as, for example, separability of the joint probability, i.e. $P(X,Y) = P(X)P(Y)$, can only be satisfied approximately in practice. Here we will be interested primarily in the degree to which the finite sample resolution affects numerical satisfiability of independence criteria expressed in terms of information-theoretic entropy $H$, e.g. $H(X|Y) = H(X)$.

For a finite sample of size $N$ the probability of the smallest non-zero event is $p_{\min} = 1/N$. This probability is also the smallest observable difference between the probabilities of any two events. Hence, for any two events the difference in their probabilities below the resolution limit $r = p_{\min}$ will be undetectable, rendering the probabilities of these observations equivalent. Conversely, the probability evaluations that produce magnitudes falling below $r$ are meaningless and can be considered noise or numerical error, at least from the data-centric perspective of the information contained in the sample. For the time being, we will assume the sample size to be ample enough for other sources of error to be negligible.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2022 doi:10.20944/preprints202202.0254.v1

Suppose $\boldsymbol{h}$ is a small perturbation of some simplex element $\boldsymbol{p}$, such that $\boldsymbol{h} \cdot \mathbf{1} = 0$, so that $\boldsymbol{p} + \boldsymbol{h}$ is again an element of the same simplex. Since entropy is a continuously differential function of its argument, the approximation of the entropy function by its Taylor expansion is

$$H(\boldsymbol{p}+\boldsymbol{h}) \approx H(\boldsymbol{p}) + \nabla H(\boldsymbol{p}) \cdot \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^T \cdot D^2 H(\boldsymbol{p}) \cdot \boldsymbol{h} + R(\boldsymbol{p}, \boldsymbol{h})$$

At the $r = p_{\min}$ resolution the smallest acceptable perturbation $\boldsymbol{h}$ must have $h_n = r$ as its n-th component and $h_m = -r$ as its m-th component, with all other components being identically zero, e.g. $\boldsymbol{h} = (0, \ldots, 0, r, 0, \ldots, 0, -r, 0, \ldots, 0)$. Evaluating the second term of the expansion gives

$$\begin{aligned} (1) \qquad \nabla H(\boldsymbol{p}) \cdot \boldsymbol{h} &= -\sum_k h_k(\log(p_k) + 1) \\ &= -r(\log(p_n) - \log(p_m)) = -r \log(p_n/p_m) \geq -r \log((1-r)/r) \end{aligned}$$

because $\sum_k h_k = 0$, and the maximum value of $\log(p_i/p_j)$ at the prescribed resolution is achieved when $p_j = r$ and $p_i = 1 - r$. In the same spirit, the third term of the expansion evaluates to

$$\frac{1}{2}\boldsymbol{h}^T \cdot D^2 H(\boldsymbol{p}) \cdot \boldsymbol{h} = -\frac{1}{2}\sum_k \frac{h_k^2}{p_k} = -\frac{r^2}{2}(1/p_n + 1/p_m) \geq -\frac{r^2}{2}(1/r + 1/r)) = -r$$

Higher order terms $R_k$ of the expansion, that comprise the residual $R(\boldsymbol{p}, \boldsymbol{h}) = \sum_k R_k(\boldsymbol{p}, \boldsymbol{h})$, present the following pattern

$$R_1(\boldsymbol{p}, \boldsymbol{h}) = \frac{1}{3!}\sum \frac{\partial^3 H(\boldsymbol{p})}{\partial p_i \partial p_j \partial p_k} h_i h_j h_k = -\frac{1}{3!}\sum \frac{-h_i^3}{p_i^2} = \frac{1}{3!} r^3 (1/p_n^2 - 1/p_m^2)$$

$$R_2(\boldsymbol{p}, \boldsymbol{h}) = \frac{1}{4!}\sum \frac{\partial^4 H(\boldsymbol{p})}{\partial p_i \partial p_j \partial p_k \partial p_l} h_i h_j h_k h_l = -\frac{1}{4!}\sum \frac{2h_i^4}{p_i^3} = -\frac{2r^4}{4!}(1/p_n^3 + 1/p_m^3)$$

$$R_3(\boldsymbol{p}, \boldsymbol{h}) = -\frac{1}{5!}\sum \frac{-2 \cdot 3 \cdot h_i^5}{p_i^4} = \frac{3! r^5}{5!}(1/p_n^4 - 1/p_m^4)$$

$$R_4(\boldsymbol{p}, \boldsymbol{h}) = -\frac{1}{6!}\sum \frac{3! \cdot 4 \cdot h_i^6}{p_i^5} = -\frac{4! r^6}{6!}(1/p_n^5 + 1/p_m^5)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2022 doi:10.20944/preprints202202.0254.v1

The series $R$ of residual terms is then bounded below by

$$R^- = \sum_{k=1}^{\infty} \frac{-2 \cdot k! \cdot r^{k+2}}{(k+2)! \cdot r^{k+1}} = -2r \sum_{k=1}^{\infty} \frac{1}{(k+1)(k+2)} = -2r/2 = -r$$

and bounded above by

$$R^+ = \sum_{k=1}^{\infty} \frac{(2k-1)! \cdot r^{2k+1}}{(2k+1)! \cdot r^{2k}} = r \sum_{k=1}^{\infty} \frac{1}{(2k)(2k+1)} \leq r \sum \frac{1}{(2k+1)^2} = r(\pi^2/8 - 1) \leq r$$

and is, therefore, convergent.

Now it is possible to assess the effect of the limited resolution on the entropy in the context of near conditional independence.

$$\begin{aligned} (2) \qquad & H(X|Y) - H(X) = \sum P(Y = y_k) H(X|Y = y_k) - H(X) \\ & \approx \sum P(Y = y_k)(H(X) + \nabla H(X) \cdot \boldsymbol{h}_k + \frac{1}{2} \boldsymbol{h}_k^T \cdot D^2 H(X) \cdot \boldsymbol{h}_k + R(X, \boldsymbol{h}_k)) - H(X) \\ & = \sum P(Y = y_k) \left( \nabla H(X) \cdot \boldsymbol{h}_k + \frac{1}{2} \boldsymbol{h}_k^T \cdot D^2 H(X) \cdot \boldsymbol{h}_k + R(X, \boldsymbol{h}_k) \right) \end{aligned}$$

A valid bound for the deviation under the circumstances of near-independence can then be estimated as

$$\begin{aligned} (3) \qquad \Delta H = |H(X|Y) - H(X)| & \leq \left| -\sum_k (r_k \log((1 - r_k)/r_k) + r_k + r_k) P(Y = y_k) \right| \\ & \leq \sum_k (\log(N_k) + 2)/N \end{aligned}$$

which, in terms of the sample size $N$, acquires the form

$$\Delta H = \sum_k (\log(N_k) + 2)/N$$

where $r_k = 1/N_k$ is the resolution limit in the set of observations conditioned on $(Y = y_k)$, and $P(Y = y_k) = N_k/N$. Although it is clearly possible, we will not concern ourselves with obtaining a tighter bound for now.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2022 doi:10.20944/preprints202202.0254.v1

Identical reasoning applies in the situation with several conditioning variables

$$
\begin{aligned}
(4) \qquad |H(X|Y,Z) - H(X|Y)| &= \left| \sum_{j,i} (P(Y=y_i, Z=z_j) H(X|Y=y_i, Z=z_j) - H(X|Y=y_i)) \right| \\
&\approx \left| \sum_{i,j} P(Y=y_i, Z=z_j) \left( \nabla H(X) \cdot \boldsymbol{h}_{ij} + \frac{1}{2} \boldsymbol{h}_{ij}^T \cdot D^2 H(X) \cdot \boldsymbol{h}_{ij} + R(X, \boldsymbol{h}_i j) \right) \right| \\
&\leq \sum_{ij} (\log(N_{ij}) + 2)/N
\end{aligned}
$$

where the resolution limit $1/N_{ij}$ corresponds to the joint event with the probability

$$P(Y=y_i, Z=z_j) = N_{ij}/N.$$

Having obtained these bounds we can now restate the local scoring function update as follows

$$\Delta S(X_i) = H(X_i|\pi_i) - H(X_i|\pi_i \cap Y) - \sum_k (\log(N_k) + 2)/N$$

where $N_k$ is now the sample count associated with the k-th state of $\pi_i \cap Y$. Note that despite our attempt to obtain sample-independent score we have retained $N$ in the second term, but this is no longer a problem because the whole term simply offsets entropy to make sure that near-independence doesn't get in the way and corresponds numerical resolution rather than the sample size. In fact, one should consider the second term to be the result of numerical or truncation error in the evaluation of conditional entropy itself, and no additional interpretation should be attributed to it. We expect to be able to perform similar analysis for other information-theoretic quantities, i.e. Mutual Information (MI), and, more importantly, Variation of Information (VI), described in [14]. More specifically, we hope to be able to modify the scoring criterion to work on the absolute scale, while retaining the near-independence behavior of the score modification reached in this paper.

## Results: Numerical verification and the sensitivity profile.

For a set of 128 uniformly distributed independent categorical variables with 8 categories and the sample size $N = 1000$ we obtain the result summarized in Table 1. For brevity, we only include 5 out of $127 \times 128 = 16256$ pairwise comparisons, which, nevertheless, is sufficiently representative, given that the statistical behavior across all pairwise comparisons will be summarized further. The negative sign

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2022 doi:10.20944/preprints202202.0254.v1

is retained in the table for ease of assessment and interpretation of the role that the penalty terms play in the evaluation of both

$$\Delta MDL = \Delta H - \Delta C$$

and

$$\Delta S = \Delta H - \Delta \xi$$

where $\Delta C$ and $\Delta \xi$ represent the MDL complexity/penalty and the resolution penalty, obtained in this work, respectively. More importantly, in the last two columns the negative sign is an indicator that the update should be rejected due to near-independence in the case of $\Delta S$, and due high storage requirements in the case of $\Delta MDL$. As we shall see further, the update rejection, equivalent to the detection of near-independence within the framework developed in this paper, isn't guaranteed for all independent variables, at least for the MDL score (see Table 4).

| $\Delta H$ | $-\Delta \xi$ | $-\Delta C$ | $\Delta S$ | $\Delta MDL$ |
|---|---|---|---|---|
| 0.02660671 | -0.05460712 | -0.16924 | -0.02800042 | -0.1426333 |
| 0.02195581 | -0.05458575 | -0.16924 | -0.03262993 | -0.14728419 |
| 0.02923805 | -0.05460464 | -0.16924 | -0.0253666 | -0.14000196 |
| 0.02532776 | -0.0546174 | -0.16924 | -0.02928964 | -0.14391225 |
| 0.03235557 | -0.05460769 | -0.16924 | -0.02225212 | -0.13688444 |

TABLE 1. 1st column is the conditional entropy deviation $\Delta H = H(X|Y) - H(X)$; 2nd column is the corresponding resolution penalty, obtained in this work; 3rd column is the MDL complexity; 4th column is the update of the resolution penalized score; 5th column is the update of MDL.

Table 2 reflects the behavior across the same 128 variables for all the 16256 pairwise comparisons. Note that the $\Delta C$ is constant, while $\Delta S$ reacts to the local properties of every pair of variables in consideration, and is a tighter bound for the deviation from independence, given by $\Delta H$.

Table 3 summarizes the results of all 16256 pairwise comparisons for 128 uniformly distributed independent variables with only 4 categories and $N = 1000$. Observe that the lower category count resulted in the decrease in the deviation from perfect independence, and this effect is also accounted by the drop in both penalty terms, although at vastly different rates.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2022 doi:10.20944/preprints202202.0254.v1

| | $\Delta H$ | $-\Delta\xi$ | $-\Delta C$ | $\Delta S$ | $\Delta MDL$ |
|---|---|---|---|---|---|
| mean | 0.02488762 | -0.05459814 | -0.16924 | -0.02971052 | -0.14435238 |
| median | 0.02453096 | -0.05460088 | -0.16924 | -0.03007069 | -0.14470904 |
| $\sigma$ | 0.00503476 | 0.00001582 | 0. | 0.00503485 | 0.00503476 |
| max | 0.04754371 | -0.05452659 | -0.16924 | -0.00704128 | -0.1216963 |

TABLE 2. Statistical summary of all 16256 pairwise comparisons of 128 independent variables with 8 categories and N=1000.

| | $\Delta H$ | $-\Delta\xi$ | $-\Delta C$ | $\Delta S$ | $\Delta MDL$ |
|---|---|---|---|---|---|
| mean | 0.00452659 | -0.03007983 | -0.0310849 | -0.02555324 | -0.02655831 |
| median | 0.00421128 | -0.03008065 | -0.0310849 | -0.02586864 | -0.02687362 |
| $\sigma$ | 0.00210545 | 0.00000414 | 0. | 0.00210548 | 0.00210545 |
| max | 0.02037829 | -0.030064 | -0.0310849 | -0.0097012 | -0.01070661 |

TABLE 3. Statistical summary of all 16256 pairwise comparisons of 128 independent variables with 4 categories and N=1000.

Table 4, on the other hand, indicates a failure of MDL to properly detect near-independent pair, as can be seen in the last row of the $\Delta MDL$ column. Here, the pairwise comparisons are carried out for 128 uniformly distributed independent variables with 2 categories and $N = 1000$, and MDL misclassifies 152 independent pairs out of 16256, approximately 1%.

| | $\Delta H$ | $-\Delta\xi$ | $-\Delta C$ | $\Delta S$ | $\Delta MDL$ |
|---|---|---|---|---|---|
| mean | 0.00048464 | -0.01642826 | -0.00345388 | -0.01594362 | -0.00296924 |
| median | 0.0002164 | -0.01642873 | -0.00345388 | -0.01621055 | -0.00323748 |
| $\sigma$ | 0.00066937 | 0.00000156 | 0. | 0.00066936 | 0.00066937 |
| max | 0.00668885 | -0.01641704 | -0.00345388 | -0.00973876 | 0.00323497 |

TABLE 4. Statistical summary of all 16256 pairwise comparisons of 128 independent variables with 2 categories and $N = 1000$.

Note that the failure of $\Delta MDL$ to identify independent variables is not resolved by an order of magnitude increase in the sample size, i.e. $N = 10000$, as can be observed in Table 5. But the total number of misclassifications of independent pairs falls to 22, which is approximately 0.14%. These observations are consistent with the known peculiarity that the MDL complexity term tends to underpenalize low

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2022 doi:10.20944/preprints202202.0254.v1

category count variable pairs, causing at times severe overfitting in the context of Bayesian Network recovery from data. Observed sample size dependence in MDL's ability to classify independent variables correctly, however, fails to explain why $\Delta C$ needs to be sensitive to sample size, given the relatively well-behaved, well-represented profile of conditional and joint events of the scenario presented here. Clearly, this undesirable underpenalizing property of MDL complexity term cannot be easily dismissed as the shortcoming of the data (see Figure 1), particularly, given the fact that $\Delta C$ depends only on $N$ and reflects nothing else about the variable pairs in question.

| | $\Delta H$ | $-\Delta\xi$ | $-\Delta C$ | $\Delta S$ | $\Delta MDL$ |
|---|---|---|---|---|---|
| mean | 0.00004935 | -0.00210343 | -0.00046052 | -0.00205408 | -0.00041117 |
| median | 0.00002315 | -0.00210343 | -0.00046052 | -0.00208029 | -0.00043737 |
| $\sigma$ | 0.00006918 | 0.00000002 | 0. | 0.00006918 | 0.00006918 |
| max | 0.00066301 | -0.00210334 | -0.00046052 | -0.00144042 | 0.0002025 |

TABLE 5. Statistical summary of all 16256 pairwise comparisons of 128 independent variables with 2 categories, $N = 10000$

To investigate this misclassification further we consider batches of 10000 randomly generated pairs of binary variables for a range of sample sizes. For every batch we extract the pair that gives the maximum value of $\Delta H$, and evaluate the corresponding values of $\Delta C$ and $\Delta\xi$. These values are then plotted against the increasing sample size in Figure 1. The MDL penalty term $\Delta C$ clearly fails to bound the deviation from independence $\Delta H$ across the whole range of sample sizes.

In Figure 2 the range of sample sizes is extended to 50000 with a coarser increment to show that the misclassification rate of $\Delta C$ sees general improvement as $N$ increases, although at $N = 46000$ the MDL penalty term once again fails to identify an independent pair. As expected, $\Delta\xi$ has no trouble in this range of parameters, and its stricter penalization profile is justified by the general volatility exhibited by $\Delta H$.

To continue, we return to our previous setup with 128 random variables and consider the scenario with 4-variate pairs for $N = 10000$. Table 6 reveals the behavior consistent with the expectations, where both updates identify near-independent pairs equally well. In this range of data parameters the terms are very close in their magnitude, so it's not surprising that the behavior is almost identical.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2022 doi:10.20944/preprints202202.0254.v1

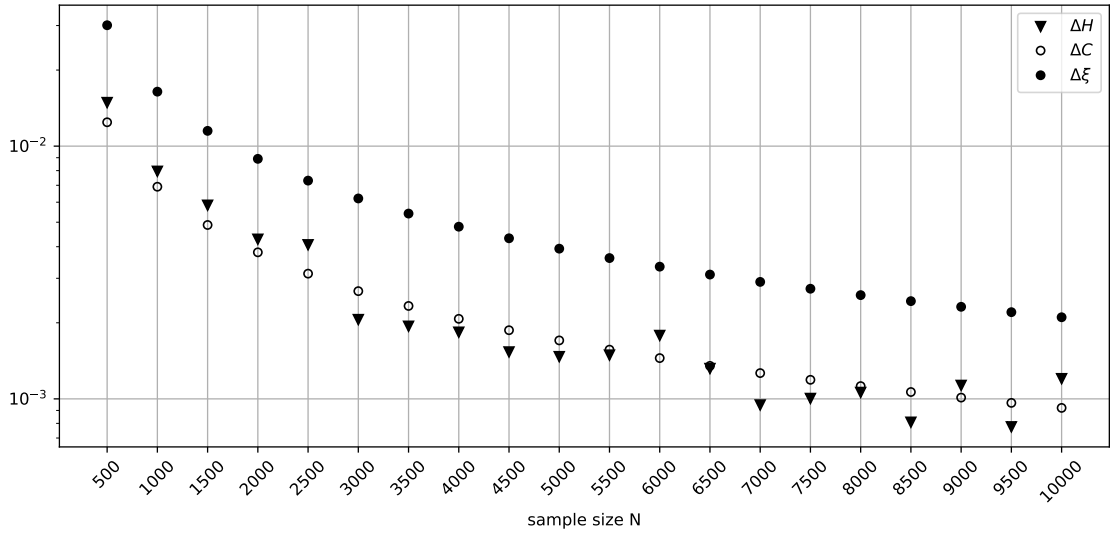


FIGURE 1. The behavior of the deviation from independence $\Delta H$, the MDL penalty term $\Delta C$ and the resolution limit $\Delta\xi$ for random binary independent variable pairs across varying sample size.

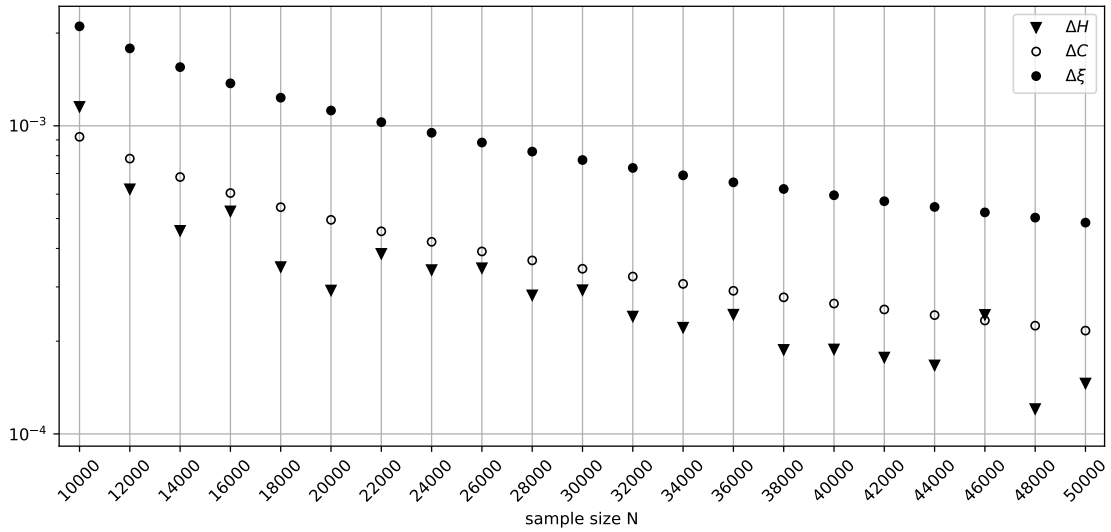


FIGURE 2. The behavior of $\Delta H$, $\Delta C$ and $\Delta\xi$ for random binary independent variable pairs across the extended sample size range.

In Table 7 the MDL penalty term is on average several times greater than $\Delta\xi$. Both scores, however, perform equally well in this range of parameters, identifying all independent pairs correctly. In this case the study is carried over the variables with 8 categories and $N = 10000$.

Table 8 reveals a misclassification on the part of $\Delta S$, as can be seen in the last row of the $\Delta S$ column. Further investigation reveals approximately 2.6% of misclassified pairs and sample size dependence of the misclassification rate which is completely resolved by increasing the samples size by one order of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2022 doi:10.20944/preprints202202.0254.v1

| | $\Delta H$ | $-\Delta\xi$ | $-\Delta C$ | $\Delta S$ | $\Delta MDL$ |
|---|---|---|---|---|---|
| mean | 0.00045076 | -0.00392956 | -0.00414465 | -0.00347879 | -0.00369389 |
| median | 0.00041775 | -0.00392957 | -0.00414465 | -0.0035118 | -0.0037269 |
| $\sigma$ | 0.00021376 | 0.00000005 | 0. | 0.00021376 | 0.00021376 |
| max | 0.00185537 | -0.00392936 | -0.00414465 | -0.00207418 | -0.00228928 |

TABLE 6. Statistical summary of all 16256 pairwise comparisons of 128 independent variables with 4 categories and $N = 10000$

| | $\Delta H$ | $-\Delta\xi$ | $-\Delta C$ | $\Delta S$ | $\Delta MDL$ |
|---|---|---|---|---|---|
| mean | 0.00245373 | -0.00730442 | -0.02256533 | -0.00485069 | -0.0201116 |
| median | 0.0024293 | -0.00730446 | -0.02256533 | -0.00487505 | -0.02013603 |
| $\sigma$ | 0.00049743 | 0.00000016 | 0. | 0.00049743 | 0.00049743 |
| max | 0.00453235 | -0.00730386 | -0.02256533 | -0.00277213 | -0.01803298 |

TABLE 7. Statistical summary of all 16256 pairwise comparisons of 128 independent variables with 8 categories and $N = 10000$

magnitude (see Table 9). This observation is fully consistent with the general understanding of the effect that limited sample size may have on conditional or joint events.

In the scenario presented here, 16-variate uniformly distributed variable can be expected to have unconditional events of the size $P(X = x_i) \approx 0.0625$. Therefore, any joint event will necessarily be smaller, in the order of the square of the unconditional events due to independence, i.e.

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) \approx 0.00390625$$

This corresponds to only roughly 40 samples per joint event in the case of $N = 10^4$. Clearly, for such small probabilities the sample size should be bigger to be adequately representative, otherwise the unaccounted for effects of sampling error may dominate the landscape.

It's not surprising that MDL seems insensitive to these circumstances, given how much it overpenalizes $\Delta H$ in this scenario. This comes at the cost of specificity, i.e. MDL would clearly fail to classify **dependent** pairs as such for all values of $\Delta H$ that would fall between, say, the values $\max \Delta H$ and $\Delta C$ presented in the table.

Table 9 reveals the ability of $\Delta S$ to recover its sensitivity under the assumption of sufficient sample size, as expected, since for $N = 10^5$ a joint event will correspond to roughly 400 samples. Note the very

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2022 doi:10.20944/preprints202202.0254.v1

| | $\Delta H$ | $-\Delta\xi$ | $-\Delta C$ | $\Delta S$ | $\Delta MDL$ |
|---|---|---|---|---|---|
| mean | 0.01130416 | -0.01349923 | -0.10361633 | -0.00219507 | -0.09231217 |
| median | 0.01127067 | -0.01349928 | -0.10361633 | -0.00222858 | -0.09234566 |
| $\sigma$ | 0.00107126 | 0.00000049 | 0. | 0.00107126 | 0.00107126 |
| max | 0.0157582 | -0.01349741 | -0.10361633 | 0.00225863 | -0.08785813 |

TABLE 8. Statistical summary of all 16256 pairwise comparisons of 128 independent variables with 16 categories and $N = 10000$

fine-tuned specificity/sensitivity ratio of $\Delta\xi$, while the MDL complexity term continues to overpenalize $\Delta H$ significantly even when provided data of ample size.

| | $\Delta H$ | $-\Delta\xi$ | $-\Delta C$ | $\Delta S$ | $\Delta MDL$ |
|---|---|---|---|---|---|
| mean | 0.0011239 | -0.00171844 | -0.01295204 | -0.00059454 | -0.01182814 |
| median | 0.00112087 | -0.00171844 | -0.01295204 | -0.00059757 | -0.01183117 |
| $\sigma$ | 0.00010628 | 0. | 0. | 0.00010628 | 0.00010628 |
| max | 0.0016554 | -0.00171843 | -0.01295204 | -0.00006305 | -0.01129664 |

TABLE 9. Statistical summary of all 16256 pairwise comparisons of 128 independent variables with 16 categories and $N = 100000$

In Figure 3 we repeat the misclassification analysis that was performed for binary variables above. The triangle pattern on this figure is the **maximum** deviation from independence obtained from batches of 10000 random 16-variate independent variable pairs for every value of $N$. The figure shows consistently improving classification precision of $\Delta\xi$, marked here by the unfilled square pattern for clarity, with a somewhat elevated sensitivity profile for smaller sample size, as expected due to the unaccounted for effect of sampling error. On the other hand, the excessive overpenalization imposed by $\Delta C$, clearly visible in this figure, is difficult to justify, given the abundant sample size and very consistent behavior on the part of $\Delta H$.

The overall impression communicated by the above results is that the resolution penalty term $\Delta\xi$ has an edge not only in interpetability, but also seems more balanced and consistent in its sensitivity/specificity profile, an aspect of direct relevance to practical performance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 February 2022 doi:10.20944/preprints202202.0254.v1

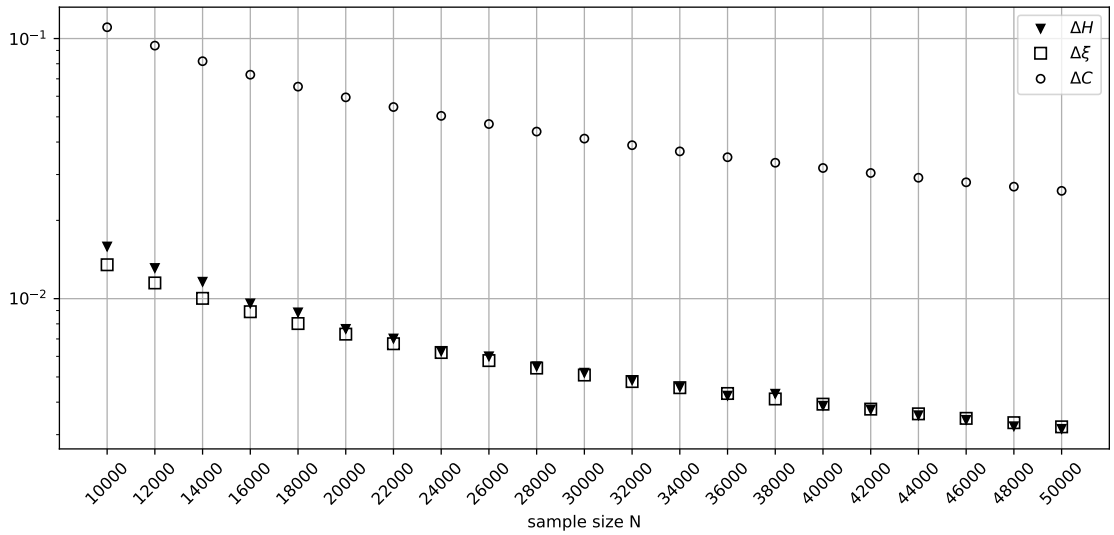


FIGURE 3. The behavior of $\Delta H$, $\Delta C$ and $\Delta \xi$ for random 16-variate independent variable pairs across varying sample sizes.

## DISCUSSION AND CONCLUSION

Numerical verification of the effect that the resolution limit has on independence assessment has shown that the reasoning presented in this work is a valid and an effective approach to managing near-independence scenarios, directly applicable in the context of data-driven recovery of Bayesian networks. The recovery method relying on the scoring criterion obtained in this paper behaves as expected, i.e. similar the method based on the MDL score, but with a slight edge in performance across the same scenarios. The advantage of the modified scoring criterion is that the numerical values obtained in the course of its application are directly interpretable in the data-source independent way, allowing for direct comparison of the results of BN reconstruction not only in robustness/stability studies, but also in scenarios where different data spans the same set of variables. Further, preliminary results indicate consistently better behavior for situations where MDL typically tends to over-fit. Hence, with this relatively simple modification we address several problems that go beyond interpretability of the score. We intend to expand this work to other information-theoretic quantities, so that the scoring function can be modified to become approximately metric and to work on an absolute scale, thus further improving the usefulness of scoring in the judgement of proximity/similarity between networks, while simultaneously addressing the numerical limitations associated with one-sided conditional entropy assessment of variable dependencies in BNs.

A number of ongoing multidisciplinary secondary biomedical data analysis studies, including (i) comparative BN analyses of multidimensional fluorescence-activated cell sorting (FACS) and other immuno-oncology datasets [1], (ii) -omics of Alzheimer's disease, (iii) BN modeling of G-protein/GPCR molecular dynamics simulation data, and (iv) BN-centered construction of gene regulatory networks from the scRNA-seq data, stimulated a significant portion of the effort detailed in this communication. Rigorous dissection of the underlying BN fundamentals and mechanics is essential for robust construction and comparison of BNs in the biomedical data analysis context.

Our experience of working with this kind of data dictates that every BN analysis in biomedical field should, in principle, allow direct comparative, possibly cross-study, investigations utilizing not only the structural, but the quantitative features of the reconstructed networks. This seemingly minor technical detail has the potential to alleviate many difficulties typically encountered when developing systematic understanding derived from the data-analytic stage of research relying on BNs.

## Acknowledgements

The authors are grateful to Arthur D. Riggs, Russell C. Rockne, Peter P. Lee and Amanda J. Myers for stimulating discussions and useful comments about the interpretability of Bayesian networks in the biomedical field. This work was partly supported by the NCI Cancer Biology System Consortium U01CA23221601 grant (to A.S.R.) and NIH NCI award P30CA033572, by the Susumu Ohno Chair in Theoretical and Computational Biology (held by A.S.R.), a Susumu Ohno Distinguished Investigator fellowship (to G.G.), and City of Hope funds (to G.G., S.B., and A.S.R.). Funding sources played no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Author Disclosure Statement

The authors declare that no competing financial interests exist.

## Data Availability Statement

Relevant code and software are available directly from the authors, or as part of the BNOmics package, at https://bitbucket.org/77D/bnomics.

## References

[1] A. S. Rodin, G. Gogoshin, S. Hilliard, L. Wang, C. Egelston, R. C. Rockne, et al., Dissecting response to cancer immunotherapy by applying bayesian network analysis to flow cytometry data, *Int. J. Mol. Sci.*, **22** (2021), 2316.

[2] X. Zhang, S. Branciamore, G. Gogoshin, A. S. Rodin, Analysis of high-resolution 3d intrachromosomal interactions aided by bayesian network modeling, *Proc. Natl. Acad. Sci. USA*, **114** (2017), E10359–E10368.

[3] S. Branciamore, G. Gogoshin, M. Di Giulio, A. S. Rodin, Intrinsic properties of TRNA molecules as deciphered via bayesian network and distribution divergence analysis, *Life (Basel)*, **8** (2018), E5

[4] G. Gogoshin, E. Boerwinkle, A. S. Rodin, New algorithm and software (bnomics) for inferring and visualizing bayesian networks from heterogeneous "big" biological and genetic data, *J. Comp. Bio.*, **24** (2017), 340–356.

[5] A. Rodin, A. Brown, A. G. Clark, C. F. Sing, E. Boerwinkle, Mining genetic epidemiology data with bayesian networks: Application to apoe gene variants and plasma lipid levels, *J. Comput. Biol.*, **12** (2005), 1–11.

[6] F. F. Sherif, N. Zayed, M. Fakhr, Discovering alzheimer genetic biomarkers using bayesian networks, *Adv. Bioinformatics*, **2015** (2015), 639367.

[7] L. Wang, P. Audenaert, T. Michoel, High-dimensional bayesian network inference from systems genetics data using genetic node ordering, *Front. Genet.*, **10** (2019), 1196.

[8] Z. Lan, Y. Zhao, J. Kang, T. Yu, Bayesian network feature finder (banff): an r package for gene network feature selection, *Bioinformatics*, **32** (2016), 3685–3687.

[9] R. Neapolitan, D. Xue, X. Jiang, Modeling the altered expression levels of genes on signaling pathways in tumors as causal bayesian networks, *Cancer Inform.*, **13** (2014), 77–84.

[10] Q. Qi, J. Li, J. Cheng, Reconstruction of metabolic pathways by combining probabilistic graphical model-based and knowledge-based methods, *BMC Proc.*, **8** (2014), S5.

[11] de Campos, Cassio and Qiang Ji., Efficient Structure Learning of Bayesian Networks using Constraints, *J. Mach. Learn. Res.*, **12** (2011): 663-689.

[12] de Campos, Luis M., A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests, *J. Mach. Learn. Res.*, **7** (2006): 2149-2187.

[13] Li, Wentian., On the Relationship between Complexity and Entropy for Markov Chains and Regular Languages, *Complex Syst.*, **5** (1991).

[14] Kraskov A. and Stögbauer H. and Andrzejak, Ralph G. and Grassberger P., Hierarchical Clustering Based on Mutual Information, arXiv:q-bio/0311039, (2003).