Discovering Reliable Correlations in Categorical Data

Panagiotis Mandros•, Mario Boley°, Jilles Vreeken⋆
•Max Planck Institute for Informatics, Saarbrücken, Germany
°Monash University, Melbourne, Australia
⋆CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
pmandros@mpi-inf.mpg.de, mario.boley@monash.edu, jv@cispa.saarland

***Abstract*—In many scientific tasks we are interested in finding correlations in our data. This raises many questions, such as how to *reliably* and *interpretably* measure correlation between a multivariate set of attributes, how to do so without having to make assumptions on data distribution or the type of correlation, and, how to search *efficiently* for the most correlated attribute sets. We answer these questions for discovery tasks with categorical data.**

In particular, we propose a corrected-for-chance, consistent, and efficient estimator for normalized total correlation, in order to obtain a reliable, interpretable, and non-parametric measure for correlation over multivariate sets. For the discovery of the top-k correlated sets, we derive an effective algorithmic framework based on a tight bounding function. This framework offers exact, approximate, and heuristic search. Empirical evaluation shows that already for small sample sizes the estimator leads to low-regret optimization outcomes, while the algorithms are shown to be highly effective for both large and high-dimensional data. Through a case study we confirm that our discovery framework identifies interesting and meaningful correlations.

***Index Terms*—knowledge discovery, information theory, total correlation, optimization, branch-and-bound**

I. Introduction

Most data are multi-dimensional, and identifying lower-dimensional correlated subsets of features is a fundamental aspect in many data analysis tasks. Such correlations are useful in many application, including the discovery of treatments for diseases, network intrusions, earthquakes etc. [1]. It is important that we can measure correlations over *multivariate* sets of features, as genes for example may reveal only a weak correlation with a disease when considered individually, while the correlation over a group of genes can be very strong [2]. It is also important that our measure is *reliable*, such that we do not discover spurious correlations, that it is *interpretable*, such that we can understand the results, and *non-parametric*, such that we do not need to assume anything about the data distribution or type of correlation. Last, but not least, as we need to be able to efficiently discover the top-k most correlated sets from possibly large quantities of data, we require an effective search framework for it.

Information theory, with the tools to quantify uncertainty, offers an attractive framework to do exactly this. We build on the concept of **total correlation**, the multivariate extension of mutual information, which non-parametrically quantifies the amount of shared information in a set of random variables [3]. Without appropriate normalization, however, scores over sets of different cardinalities are not comparable, which is a problem when searching for the top correlations [4]. We hence consider **normalized total correlation**, which does not only address this, but is also interpretable: a score of 0 corresponds to statistically independent variables, while a score of 1 to the existence of a variable that "explains" all others.

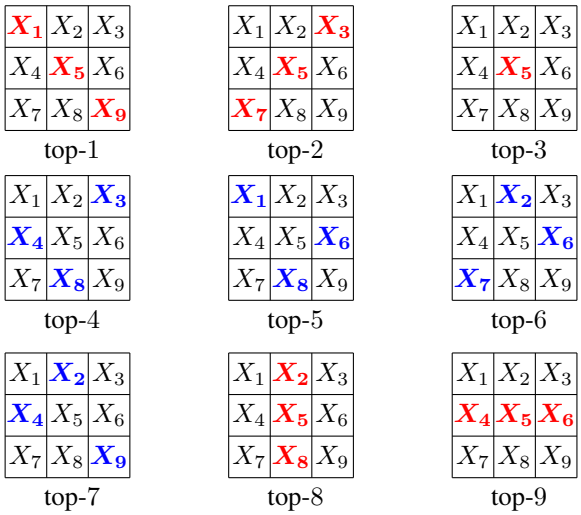

Figure 1: **Top correlated sets discovered on Tic-tac-toe**. Color indicates the selected cells, with red designating the inclusion of X_{10} that corresponds to the binary outcome of the game. In a nutshell, red and blue correlated sets can be interpreted as latent factors for win and loss, respectively. (Sec.V-C)

Although theoretically sound, in practice the score is *unreliable* when we estimate it from empirical data: due to sparsity the plug-in estimator leads to inflated estimates [5]. In our case in particular, the data sparsity induced by the increasingly larger sets of variables we have to consider during optimization, can lead to many false discoveries (see Fig. 2 for a demonstration). In addition, the normalized total correlation is difficult to optimize; it is neither monotone, nor submodular, and hence the resulting combinatorial optimization problem for discovering the top correlated sets is hard to solve efficiently.

To address each of these issues, we build upon the recent advances on deriving corrected-for-chance information-theoretic estimators well-suited for optimization [6], [7], and

propose a *reliable* and *efficient* estimator for normalized total correlation. Furthermore, we enable effective exhaustive and heuristic algorithms for the discovery of the top correlated sets by exploiting various structural properties of the estimator proposed. Experimental evaluation shows that the estimator has attractive statistical properties, the algorithms proposed are indeed effective on a wide range of benchmark data, and finally, concrete findings in real data show that our framework discovers interesting and sensible information (see Fig. 1). In summary, our main contributions are the following: we

i) propose a consistent, corrected-for-chance, and efficient estimator for the normalized total correlation (Sec. III),
ii) provide effective algorithms for exact, approximate, and heuristic search (Sec. IV), and finally
iii) perform empirical evaluation on a wide range of real and synthetic datasets (Sec. V).

We start with preliminaries and problem definition in Sec. II, and round up with a concluding discussion in Sec. VI. More details, proofs, and additional experiments, can be found in the extended version of the paper [8].

II. Problem Definition

We consider data $\mathbf{D}_n$ consisting of n i.i.d. samples from a set of d categorical random variables $\mathcal{I} = \{X_1, \ldots, X_d\}$, with joint distribution $p(X_1, \ldots, X_d)$, domains V_X, and domain sizes $S_X = |V_X|$. We are interested in discovering subsets $\mathcal{X} \subseteq \mathcal{I}$ in $\mathbf{D}_n$ that exhibit high correlation/redundancy with respect to the unsupervised information-theoretic concept of total correlation introduced by Watanabe [3].

The **total correlation** for a set of variables $\mathcal{X} = \{X_1, \ldots, X_m\}$ is defined as

$$W(\mathcal{X}) = \sum_{X \in \mathcal{X}} \Big(H(X) \Big) - H(\mathcal{X}) = \sum_{i=2}^{m} I(\mathcal{X}_{i-1}; X_i) \ ,$$

where $\mathcal{X}_i$ represents the set $\{X_j \in \mathcal{X} \colon j \leq i \leq m\}$, with $\mathcal{X}_0$ being the empty set. Here, H denotes the **Shannon entropy** [9], defined as $H(X) = -\sum_{x \in V_X} p(x) \log p(x)$ for random variable X, and quantifies its uncertainty in bits of information, assuming logarithm with base 2. Moreover, $H(X \mid Y)$ denotes the **conditional entropy** of X given another random variable Y, i.e., $H(X \mid Y) = \sum_{y \in V_Y} p(y) H(X \mid Y = y)$, and quantifies the uncertainty of X conditioned on Y. Lastly, $I(X;Y) = H(X) - H(X \mid Y)$ is the **mutual information**, and measures the amount of shared information between X and Y. Essentially, total correlation is a multivariate correlation/redundancy measure quantifying the total amount of shared information in a set of random variables. As a function of p, total correlation is order-invariant, and it holds that $W(\mathcal{X}) \geq 0$, with equality if and only if all variables $X \in \mathcal{X}$ are statistically independent.

Total correlation, however, is not suitable for comparing the degree of correlation between different sets of variables, since cardinalities, joint and marginal entropies, all vary. In addition, total correlation lacks an intuitive and intepretable scale, e.g., in $[0, 1]$. These can be resolved by expressing how far the correlation in a set of variables is from the scenario of them being maximally correlated. To achieve this, we present the following proposition.

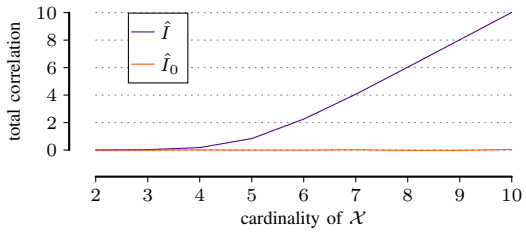

Figure 2: **Correlation-by-chance.** Estimated total correlation for variable set $\mathcal{X}$ of increasing cardinality. All variables are uniformly and independently sampled with domain size 4 and sample size 1000. Population value for total correlation is 0. Correlation increases when naive estimator $\hat{I}$ is used, but not for the corrected-for-chance $\hat{I}_0$. ([7], also defined in Sec. III)

Proposition 1. *Given a set of variables $\mathcal{X} = \{X_1, \ldots, X_m\}$, we have that*

a) $W(\mathcal{X}) \leq \sum_{X \in \mathcal{X}} H(X) - \max_{X \in \mathcal{X}} H(X)$,
b) with equality iff $\exists X_i \in \mathcal{X}$ s.t., $X_j = f(X_i), \forall X_j \in \mathcal{X}$.

Defining $\bar{W}(\mathcal{X}) = \sum_{X \in \mathcal{X}} H(X) - \max_{X \in \mathcal{X}} H(X)$, we obtain the **normalized total correlation** as

$$w(\mathcal{X}) = W(\mathcal{X}) / \bar{W}(\mathcal{X}) \ ,$$

for which it holds that $w(\mathcal{X}) \in [0, 1]$, with 0 being the case where all $X \in \mathcal{X}$ are statistically independent, and 1 when there exists a variable that "explains" all others. By quantifying the percentage of correlation within $\mathcal{X}$, the score is now better interpretable, as well as comparable across the different variable sets with varying joint and marginal entropies.

The data samples $\mathbf{D}_n$ induce an empirical distribution $\hat{p}$ defined using the empirical counts of values in $\mathbf{D}_n$, from which plug-in estimators can be derived for all the aforementioned quantities, i.e., $\hat{H}, \hat{I}, \hat{W}, \hat{w}$. These estimators, however, are known to have biases that depend on the domain sizes of the variables involved [10], with mutual information, in particular, having a positive bias. While it is easier in general to obtain good estimates for marginal quantities, total correlation involves mutual information terms that need to be estimated for increasingly larger sets of variables. This can lead to situations with severely inflated estimates (see Fig. 2 for a demonstration).

Even if a more suitable estimator was available, it remains unclear how to efficiently solve the resulting combinatorial optimization problem for finding the top correlated sets $\mathcal{X}^*$ in $\mathbf{D}_n$. Hence, in order to have an overall useful method for our task, we need to a) derive a corrected-for-chance estimator $\hat{w}'$ for w, and b) find an effective solution to the optimization problem by exploiting structural properties of $\hat{w}'$. We present solutions to these in Sec. III and Sec. IV respectively.

III. Reliable normalized total correlation

In this section we derive a corrected for chance, consistent, and efficient to compute estimator for the normalized total correlation. The estimator follows the idea of correcting the plug-in by subtracting values of null hypothesis models, leading to either parametric (e.g., [6]), or non-parametric solutions (e.g., [11]). Unlike the plug-in, such estimators give conservative estimates for sparse data in high-dimensional spaces, making them therefore well-suited for reliable optimization.

For the non-parametric case, we proposed a reliable estimator for mutual information [7] defined as

$$\hat{I}_0(X;Y) = \hat{I}(X;Y) - E_0[\hat{I}(X;Y)] \ ,$$

where $E_0[\hat{I}(\mathcal{X};Y)]$ is the expected value of $\hat{I}$ under the **permutation model** [12, p. 214], a non-parametric independence model for contingency tables that assumes fixed marginal counts. The expected value under this model is equal to $E_0[\hat{I}(\mathcal{X};Y)] = \sum_{\sigma\in S_n} \hat{I}(X;Y_\sigma)/n!$, where S_n denotes the symmetric group for n, i.e., the set of all bijections from $\{1,\dots,n\}$ to $\{1,\dots,n\}$, and Y_σ denotes the Y samples permuted according to a $\sigma \in S_n$. Exploiting symmetries, this value can be computed in $O(n\max\{S_X, S_Y\})$ (see [13], [14] for the computation, and [15] for the complexity). For the rest of this paper we denote $E_0[\hat{I}(X;Y)]$ with $m_0(X,Y,n)$.

Following the same non-parametric correction principle, and assuming we can adequately estimate marginal entropies $\hat{H}(X)$, we can define a corrected-for-chance estimator for the normalized total correlation by plugging $\hat{I}_0$ and arrive at

$$\sum_{i=2}^{m} \Big(\hat{I}(\mathcal{X}_{i-1};X_i) - m_0(\mathcal{X}_{i-1},X_i,n)\Big)/\bar{W}(\mathcal{X}) \ .$$

However, unlike the plug-in $\hat{w}$, this estimator violates the order-invariance of total correlation since the correction m_0 is not a function of $\hat{p}$, but rather a function of domain sizes and marginal counts. To ensure order-invariance, we select the order of variables that leads to the most conservative estimate for the normalized total correlation, which translates to the order that maximizes the correction term, i.e.,

$$\begin{aligned}\hat{w}_0(\mathcal{X}) =& \frac{\sum_{i=2}^{m} \hat{I}(\mathcal{X}_{i-1};X_i)}{\bar{W}(\mathcal{X})} \\ &- \frac{\max\limits_{\sigma\in S_m} \sum_{i=2}^{m} m_0(\mathcal{X}_{\sigma(i-1)},X_{\sigma(i)},n)}{\bar{W}(\mathcal{X})} \\ =& \hat{w}(\mathcal{X}) - t_0(\mathcal{X},n) \ ,\end{aligned}$$

where $\mathcal{X}_\sigma$ denotes set $\mathcal{X}$ ordered according to a $\sigma \in S_m$.

Regarding efficiency, $\hat{w}_0$ is clearly infeasible to compute in practice. For a set of m variables, there are $m-1$ calculations of the permutation model with each subsequent calculation having an increased cost (since domain sizes $S_{\mathcal{X}_{\sigma(i-1)}}$ can grow exponentially with i), and there are $m!$ possible permutations to find the maximum correction term, resulting in a total complexity of $O(m^2(m-1)!nS_{\mathcal{X}})$. We dramatically reduce this complexity by first replacing the exact calculation of the expected value m_0 with an upper bound, and then propose a relaxation to this bound such that we can efficiently find the order $\sigma^* \in S_m$ of variables maximizing the correction term.

Proposition 2 ([11], Thm. 7)**.** *For variables X, Y, with domain sizes S_X, S_Y, and sample size n, it holds that*

$$m_0(X,Y,n) \le \log \frac{n + S_X S_Y - S_X - S_Y}{n-1} \ .$$

We denote this upper bound with $m_{\bar{0}}(X,Y,n)$, and the corresponding correction term $t_{\bar{0}}(\mathcal{X},n)$, i.e.,

$$t_{\bar{0}}(\mathcal{X},n) = \max_{\sigma\in S_m} \sum_{i=2}^{m} m_{\bar{0}}(\mathcal{X}_{\sigma(i-1)},X_{\sigma(i)},n)/\bar{W}(\mathcal{X}) \ .$$

Now, while the exact expected values have been replaced with something more efficient, $t_{\bar{0}}(\mathcal{X},n)$ as function of the joint domain sizes $S_{\mathcal{X}_{\sigma(i-1)}}$ remains infeasible: for every $\sigma \in S_m$ and $i \in [2,m]$, we need to compute the joint domain size of $\mathcal{X}_{\sigma(i-1)}$ with $X_{\sigma(i)}$. We proceed to relax this requirement.

Assuming a strictly positive distribution p, i.e., $p(\mathcal{X}=\mathbf{x}) > 0$ for all $\mathcal{X} \subseteq \mathcal{I}$ and $\mathbf{x} \in V_{\mathcal{X}}$, joint domain sizes can be written as a product of marginal domain sizes, i.e., $S_{\mathcal{X}} = \prod_{X\in\mathcal{X}} S_X$. Furthermore, a relaxation that considers only the joint contribution of the variables in $\mathcal{X}$, leads to the bound

$$m_{\bar{\bar{0}}}(\mathcal{X}_{i-1},X_i,n) = \log \frac{n + \big(\prod_{X\in\mathcal{X}_{i-1}} S_X\big) S_{X_i}}{n-1} \ ,$$

and to the following correction term

$$t_{\bar{\bar{0}}}(\mathcal{X},n) = \max_{\sigma\in S_m} \sum_{i=2}^{m} m_{\bar{\bar{0}}}(\mathcal{X}_{\sigma(i-1)},X_{\sigma(i)},n)/\bar{W}(\mathcal{X}) \ .$$

In the following theorem we establish that this quantity is both a consistent upper bound for $t_{\bar{\bar{0}}}$, and efficient to compute without explicitly considering all permutations $\sigma \in S_m$.

Theorem 1. *For set of variables $\mathcal{X} = \{X_1,\dots,X_m\}$, it holds*

a) $t_{\bar{\bar{0}}}(\mathcal{X},n) \ge t_{\bar{0}}(\mathcal{X},n)$
b) $\lim_{n\to\infty} t_{\bar{\bar{0}}}(\mathcal{X},n) = 0$
c) $\sum_{i=2}^{m} m_{\bar{\bar{0}}}(\mathcal{X}_{\sigma(i-1)},X_{\sigma(i)},n)$ *is maximized for* $\sigma^* \in S_m$ *with* $S_{X_{\sigma^*(1)}} \ge S_{X_{\sigma^*(2)}} \cdots \ge S_{X_{\sigma^*(m)}}$

We now have an efficiently computable correction term $t_{\bar{\bar{0}}}(\mathcal{X},n)$, going from an initial complexity of $O(m^2(m-1)!nS_{\mathcal{X}})$, to that of $O(m + m\log m)$, where $m\log m$ is for sorting the domain sizes S_X, for $X \in \mathcal{X}$. In addition, as an upper bound to $t_{\bar{0}}$, this correction is as conservative with regards to its estimates, which is a design goal for reliability. Finally, we arrive at the **reliable normalized total correlation**

$$\hat{w}_{\bar{\bar{0}}}(\mathcal{X}) = \hat{w}(\mathcal{X}) - t_{\bar{\bar{0}}}(\mathcal{X},n) \ .$$

In addition to being very efficient, the consistency of the plug-in $\hat{H}$ (see, e.g., [16]), together with Th. 1b), implies that $\hat{w}_{\bar{\bar{0}}}$ is a consistent estimator for the normalized total correlation.

IV. Optimization

Here, we provide algorithms for the following optimization problem: given data D_n consisting of n i.i.d. samples of random variables $\mathcal{I} = \{X_1, \ldots, X_d\}$, as well as a positive integer k, find the top-k subsets $\mathcal{X}_1^*, \ldots, \mathcal{X}_k^* \subseteq \mathcal{I}$ with

$$\hat{w}_{\bar{\emptyset}}(\mathcal{X}_i^*) = \max\{\hat{w}_{\bar{\emptyset}}(\mathcal{X}) \colon \hat{w}_{\bar{\emptyset}}(\mathcal{X}_{i-1}^*) \geq \hat{w}_{\bar{\emptyset}}(\mathcal{X}), \mathcal{X} \subseteq \mathcal{I}\}. \quad (1)$$

As is common in hard combinatorial problems, we instantiate the **branch-and-bound** framework to obtain an exact algorithm for Eq (1). This framework consists of two main ingredients: a branch operator to enumerate some abstract search space Ω, and an admissible bounding function for the optimization function $f \colon \Omega \to \mathbb{R}$ at hand. The **branch operator** is a function $\mathbf{r} \colon \mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$ that non-redundantly generates the search space from some designated root element $\perp \in \Omega$, i.e., for all $\omega \in \Omega$ there must be a unique sequence $\perp = \omega_1, \ldots, \omega_l = \omega$ such that $\omega_{i+1} \in \mathbf{r}(\omega_i)$ for $i = 1, \ldots, l-1$.

An **admissible bounding function** $\bar{f}$, also known as optimistic estimator, must guarantee the property $\bar{f}(\omega) \geq \max\{f(\omega') \colon \omega' \in \mathbf{r}^*(\omega)\}$, where $\mathbf{r}^*(\omega)$ denotes the set of all $\omega' \in \Omega$ that can be generated from ω by multiple applications of $\mathbf{r}$. The value $\bar{f}(\omega)$ is called the **potential** of element ω. With these, a branch-and-bound algorithm enumerates Ω starting from $\perp$, tracks the best solution, and prunes expanding elements with $\bar{f}$ that cannot yield an improvement over the best solution. In addition, the framework provides the option of relaxing the required result guarantee to that of an α**-approximation** for accuracy parameter $\alpha \in (0, 1]$. Therefore, an $\alpha < 1$ allows to trade accuracy for efficiency in a principled manner.

The ideal bounding function for $\hat{w}_{\bar{\emptyset}}$ in our case would be

$$\bar{w}_{\bar{\emptyset}}^*(\mathcal{X}) = \max\{\hat{w}_{\bar{\emptyset}}(\mathcal{X}') \colon \mathcal{X} \subseteq \mathcal{X}' \subseteq \mathcal{I}\} \ .$$

Efficiently computing this function, however, would imply an efficient algorithm for the original optimization problem. Instead, we shift our attention into independently deriving tight bounds for the two terms of $\hat{w}_{\bar{\emptyset}}(\mathcal{X})$, i.e., an upper bound for $\hat{w}(\mathcal{X})$ and a lower bound for $t_{\bar{\emptyset}}(\mathcal{X}, n)$, in order to arrive at a looser, but efficient to compute bounding function. Here, however, it is not possible to both derive tight bounds and also guarantee their admissibility for arbitrarily enumerated search spaces. The difficulty stems from the inability to "predict" their behavior with respect to the subset relation—both numerators are monotonically increasing functions, but this property does not extend together with the normalizer $\bar{W}(\mathcal{X})$. For example, for a $\mathcal{X}' \supseteq \mathcal{X}$ it might be that $t_{\bar{\emptyset}}(\mathcal{X}', n) \geq t_{\bar{\emptyset}}(\mathcal{X}, n)$, but for a different superset $\mathcal{X}'' \supseteq \mathcal{X}$ that $t_{\bar{\emptyset}}(\mathcal{X}'', n) \leq t_{\bar{\emptyset}}(\mathcal{X}, n)$.

It turns out, that under a more strict partial order we can induce a certain structure into our problem that allow us to derive tight, admissible bounds for both terms.

Definition 1. Given $\mathcal{I} = \{X_1, \ldots, X_d\}$, we say that $\mathcal{X}' \subseteq \mathcal{I}$ is a **low entropy extension** of a $\mathcal{X} \subseteq \mathcal{I}$, denoted as $\mathcal{X} \subseteq_H \mathcal{X}'$, if $\mathcal{X} \subseteq \mathcal{X}'$, and for all $X' \in \mathcal{X}' \setminus \mathcal{X}$, $\hat{H}(X') \leq \min_{X \in \mathcal{X}} \hat{H}(X)$.

We can guarantee that this partial order holds in the enumerated search space by simply considering a decreasing-entropy branching operator of the form

$$\mathbf{r}_H(\mathcal{X}) = \{\mathcal{X} \cup \{X\} \colon \hat{H}(X) \leq \min_{X' \in \mathcal{X}} \hat{H}(X'), X \in \mathcal{I} \setminus \mathcal{X}\} \ ,$$

i.e., it holds that $\mathcal{X} \subseteq_H \mathcal{X}'$ for all $\mathcal{X}' \in \mathbf{r}_H(\mathcal{X})$. This operator is equivalent to the standard alphabetical enumeration order, i.e., $\mathbf{r}_A(\mathcal{X}) = \{\mathcal{X} \cup \{X_i\} \colon i > \max\{j \colon X_j \in \mathcal{X}\}, i \leq d\}$, after initially sorting $\mathcal{I}$ in descending entropy order. We now proceed with showing that under this partial order, the correction term $t_{\bar{\emptyset}}$ is monotonically increasing.

Theorem 2. *For subsets $\mathcal{X}, \mathcal{X}'$ of $\mathcal{I}$ with $\mathcal{X} \subseteq_H \mathcal{X}'$, it holds that $t_{\bar{\emptyset}}(\mathcal{X}, n) \leq t_{\bar{\emptyset}}(\mathcal{X}', n)$.*

Following from the theorem, a **trivial bounding function** can be derived using the upper bound 1 for $\hat{w}(\mathcal{X})$, i.e.,

$$\begin{aligned}\hat{w}_{\bar{\emptyset}}(\mathcal{X}') =& \hat{w}(\mathcal{X}') - t_{\bar{\emptyset}}(\mathcal{X}', n) \\ \leq& 1 - t_{\bar{\emptyset}}(\mathcal{X}, n) = \bar{w}_{\bar{\emptyset}\text{mon}}(\mathcal{X}) \ ,\end{aligned}$$

for all $\mathcal{X}'$ that are low entropy extensions of $\mathcal{X}$. It is clear, however, that $\bar{w}_{\bar{\emptyset}\text{mon}}(\mathcal{X})$ is not tight: it upper bounds $\hat{w}(\mathcal{X})$ with the maximum possible value for the normalized total correlation, without taking into consideration both the correlation in $\mathcal{X}$ so far, nor any information with regards to the remaining branch. We derive a much tighter upper bound for $\hat{w}$ by further exploiting the structure of the branch operator. We define $R_{\mathcal{X}} = \{X \in \mathcal{I} \setminus \mathcal{X} \colon \hat{H}(X) \leq \min_{X' \in \mathcal{X}} \hat{H}(X')\}$ to be the set of all refinement elements of $\mathcal{X}$, and $\bar{w}(\mathcal{X})$ the quantity

$$\bar{w}(\mathcal{X}) = \frac{\sum_{i=2}^{m} \hat{I}(\mathcal{X}_{i-1}; X_i) + \sum_{X' \in R_{\mathcal{X}}} \hat{H}(X')}{\bar{W}(\mathcal{X}) + \sum_{X' \in R_{\mathcal{X}}} \hat{H}(X')} \ ,$$

i.e., the plug-in $\hat{w}(\mathcal{X})$ after adding the marginal entropies of the refinement elements of $\mathcal{X}$. The following theorem establishes that $\bar{w}(\mathcal{X})$ is an upper bound to $\hat{w}(\mathcal{X})$ with respect to $\subseteq_H$.

Theorem 3. *For a $\mathcal{X} \subseteq \mathcal{I}$ and any $\mathcal{X}' \subseteq \mathcal{I}$ with $\mathcal{X} \subseteq_H \mathcal{X}'$, it holds that $\bar{w}(\mathcal{X}) \geq \hat{w}(\mathcal{X}')$.*

We can now define the **tighter bounding function** $\bar{w}_{\bar{\emptyset}\text{ref}}(\mathcal{X}) = \bar{w}(\mathcal{X}) - t_{\bar{\emptyset}}(\mathcal{X}, n)$, which has an extra $O(|\mathcal{R}_{\mathcal{X}}|)$ complexity compared to $\bar{w}_{\bar{\emptyset}\text{mon}}(\mathcal{X})$. Note that in practice we use both: first evaluate $\bar{w}_{\bar{\emptyset}\text{mon}}$ that we get for free by caching $t_{\bar{\emptyset}}$ after computing $\hat{w}_{\bar{\emptyset}}$, and then proceed with $\bar{w}_{\bar{\emptyset}\text{ref}}$ if it fails.

We summarize the resulting **exhaustive** method for the discovery of reliable correlated sets in Algorithm 1 in [8]. For heuristic search, we consider the standard **greedy** algorithm, i.e., level-wise search where only the best candidate is refined, coupled with $\mathbf{r}_H$ and $\bar{w}_{\bar{\emptyset}\text{ref}}$ for pruning.

V. Evaluation

In this section we empirically evaluate the proposed correlation discovery framework. We perform experiments on synthetic data in order to investigate the performance of the estimators, we use a wide selection of benchmark data to evaluate the performance of the algorithms and bounding function $\bar{w}_{\bar{\emptyset}}$, as well as provide concrete findings in example exploratory tasks.

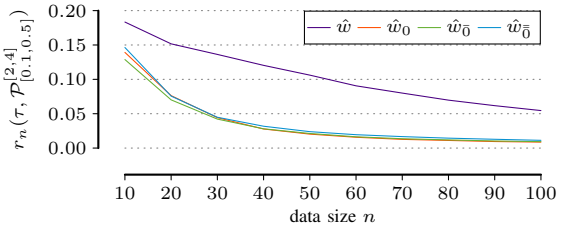

Figure 3: **Average regret.** Regret $r_n(\tau, \mathcal{P}^{[2,4]}_{[0.1,0.5]})$ for sample sizes $n = \{10, \ldots, 100\}$ and estimators $\tau = \{\hat{w}, \hat{w}_0, \hat{w}_{\bar{0}}, \hat{w}_{\bar{\bar{0}}}\}$.

A. Estimator performance

Here we evaluate the performance of the corrected-for-chance estimators $\hat{w}_0, \hat{w}_{\bar{0}}, \hat{w}_{\bar{\bar{0}}}$ proposed and the plug-in $\hat{w}$. For this evaluation, we create synthetic data in the following way. We randomly and uniformly sample joint probability distributions $p^{(i)} \in \mathcal{P}^d_{[a,b]}$, where $\mathcal{P}^d_{[a,b]}$ denotes the set of all joint probability distributions with d dependent random variables and resulting w score in $[a, b]$. Each random variable has domain size 3. For example, $\mathcal{P}^4_{[0,0.3]}$ is the set of probability distributions $p(\mathcal{X})$, $\mathcal{X} = \{X_1, \ldots, X_4\}$, with $S_{X_i} = 3$, and $w(\mathcal{X}) \in [0, 0.3]$. We augment these distributions with 3 independent and uniformly distributed random variables, also of domain size 3. Each $p^{(i)} \in \mathcal{P}^d_{[a,b]}$ has then its own set of $2^{d+3} - 1$ marginalized distributions for which we can compute the w score.

We consider dimensionalities $d = 2, 3, 4$, and four different regimes $P^d_{[0.1,0.2)}$, $P^d_{[0.2,0.3)}$, $P^d_{[0.3,0.4)}$, $P^d_{[0.4,0.5]}$, representing weak, low, medium, and high correlation. We sample one distribution for each combination, resulting in 12 different distributions $p^{(i)}, i = 1, \ldots, 12$. We consider data sizes $n = \{10, 20, 30, \ldots, 100\}$, and for each $p^{(i)}$ and n we sample 500 datasets according to $p^{(i)}$ and denote them as $\mathbf{D}^{(i)}_{n,j}, j \in [1, 500]$. We pick $n = \{10, \ldots, 100\}$, as it is expected, given that all estimators are consistent, that their behavior carries on for larger sample sizes and distributions.

We evaluate the estimators using regret, as it is an accurate summary for consistency, convergence, and generalization error. The **regret** is defined as $r_n(\tau, p^{(i)}) = \mathbb{E}\big[w(\mathcal{X}^*_i) - w(\mathcal{X}^*_{i,j,n,\tau})\big]$, where $\mathcal{X}^*_i$ represents the true maximizer of population $p^{(i)}$, and $\mathcal{X}^*_{i,j,n,\tau}$ the maximizer in $\mathbf{D}^{(i)}_{n,j}$ according to an estimator $\tau = \{\hat{w}, \hat{w}_0, \hat{w}_{\bar{0}}, \hat{w}_{\bar{\bar{0}}}\}$, for which we use exhaustive search to obtain (to compute the inefficient $\hat{w}_0, \hat{w}_{\bar{0}}$, we use the decreasing entropy permutation). The expected value is with respect to $j \in [1, 500]$. We average regrets across the different $p^{(i)}$ to obtain $r_n(\tau, \mathcal{P}^{[u,v]}_{[a,b]})$, e.g., $r_n(\tau, \mathcal{P}^{[2,3]}_{[0,0.5]})$ would be the average regret of estimator τ across all $p^{(i)} \in \mathcal{P}^3_{[0,0.5]}$ and $p^{(i)} \in \mathcal{P}^4_{[0,0.5]}$.

We start with Fig. 3 and plot $r_n(\tau, \mathcal{P}^{[2,4]}_{[0.1,0.5]})$, i.e. the average regret across all $p^{(i)}$. We observe that in general, the corrected estimators perform well. They have a smaller regret across all n, and for some n there is even a factor of 5 improvement. In addition, they converge faster to a regret close to 0. Regarding the efficient $\hat{w}_{\bar{\bar{0}}}$, we see that despite the necessary relaxations, it has performance that is on par with both $\hat{w}_0$ and $\hat{w}_{\bar{0}}$.

Finally, in Fig. 4 we plot the regrets averaged over the different dimensionalities, i.e., $r_n(\tau, \mathcal{P}^2_{[0.1,0.5]})$, $r_n(\tau, \mathcal{P}^3_{[0.1,0.5]})$, and $r_n(\tau, \mathcal{P}^4_{[0.1,0.5]})$. Under this different view, we see that the plug-in estimator $\hat{w}$ has an increasing difficulty to converge to 0 regret with respect to dimensionality, while the corrected estimators do not exhibit this behavior, as expected. Among the corrected, the differences are more profound for $d = 2$ with $\hat{w}_{\bar{\bar{0}}}$ having worse performance. Overall, we see that our proposed corrected-for-chance estimators $\hat{w}_0, \hat{w}_{\bar{0}}$, and $\hat{w}_{\bar{\bar{0}}}$, clearly outperform the plug-in, sometimes even by a factor of 5. In addition, we observe that the efficiently computable $\hat{w}_{\bar{\bar{0}}}$ has statistical properties that are on par with $\hat{w}_0$ and $\hat{w}_{\bar{0}}$.

B. Optimization performance

In this section we investigate the performance of the bounding function $\bar{w}_{\bar{\bar{0}}\text{ref}}$ and algorithms proposed for exhaustive (**BnB**) and heuristic search (**Greedy**) for the reliable normalized total correlation $\hat{w}_{\bar{\bar{0}}}$. We consider benchmark data from the KEEL data repository, and particularly all classification datasets with no missing values and $d \geq 7$, resulting in 49 datasets with $n \in [101, 1025010]$ and $d \in [7, 91]$, summarized in Table I in [8]. All metric attributes are discretized in 5 equal-frequency bins. Our code is available online.[1]

We employ the two algorithms in order to retrieve the top correlated set. For BnB, we set α to be the highest possible in increments of 0.05 such that it terminates in less than 30 minutes, and report in Table I the runtime, the percentage of the pruned search space, the depth of the solution, the maximum depth BnB had to selectively reach, and the quality $\hat{w}_{\bar{\bar{0}}}$ of the top correlated set [8]. For Greedy we report runtime and the difference of the quality for the top result with that from BnB. We average runtimes over 3 independent executions.

We observe that BnB is highly efficient as it finds the optimum solution ($\alpha = 1$) in less than 30 minutes for 42 out of 49 datasets. In 30 of them, it takes less than a minute. For all 49, it requires 77 seconds on average. The bounding function $\bar{w}_{\bar{\bar{0}}\text{ref}}$ is very effective in pruning, enabling the discovery of optimum solutions on datasets such as *coil2000* and *move. libras* with 86 and 91 attributes, that with exhaustive search would otherwise be impossible. In addition, an average of 5 maximum depth combined with an average solution size of 2.2, shows that the synergy of $\bar{w}_{\bar{\bar{0}}\text{ref}}$ and enumerated search space allows to selectively explore based on the structure of the data, and not simply by cardinality.

The Greedy algorithm requires only a couple of seconds on the majority of the datasets. On average, it terminates after 3 seconds. In addition, the solutions produced by Greedy are almost optimal considering that there are only 2 negligible cases where the two algorithms differ. In general, for a solution on the second level Greedy cannot "stray" enough. We do observe, however, that even for solution cardinalities of 3 and 4, Greedy solutions are identical to those of BnB.

Overall, both algorithms are very effective with $\bar{w}_{\bar{\bar{0}}\text{ref}}$ as a bounding function. The BnB algorithm would be preferable in

[1] https://github.com/pmandros/wodiscovery

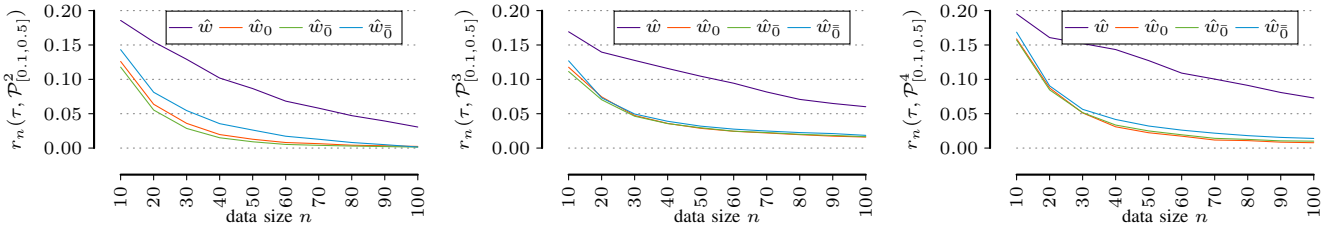

Figure 4: **Regret curves averaged over different dimensionalities.** Average regret $r_n(\tau, \mathcal{P}^2_{[0.1,0.5]})$ **(left)**, $r_n(\tau, \mathcal{P}^3_{[0.1,0.5]})$ **(middle)**, and $r_n(\tau, \mathcal{P}^4_{[0.1,0.5]})$ **(right)**, for sample sizes $n = \{10, \ldots, 100\}$ and estimators $\tau = \{\hat{w}, \hat{w}_0, \hat{w}_{\bar{0}}, \hat{w}_{\bar{\bar{0}}}\}$.

scenarios were solution guarantees are required, while GREEDY when efficiency is more important, e.g., on very large datasets.

C. Example discoveries

Last, we proceed with presenting concrete findings on **Tic-tac-toe**, a game of two players picking a symbol from $\{x, o\}$ and, taking turns, mark their symbols in an unoccupied cell of a 3×3 game board. A player wins the game with 3 consecutive cells in a row, column, or diagonal. The dataset consists of 958 end-game, winning configurations, i.e., there are no draws. There are 10 input variables $\mathcal{I} = \{X_1, \ldots, X_{10}\}$, where $X_i, i \in [1, 9]$ represent the cells of the board, taking values in $\{x, o, b\}$ with b denoting an empty cell, and X_{10} is the binary outcome of the game for player with symbol x.

We present in Fig. 1 the top-9 results retrieved with $\hat{w}_{\bar{\bar{0}}}$. The variables $X_1, \ldots, X_9$ are mapped to their corresponding board positions and color indicates the result. Red designates the result set contains X_{10}. We observe that top-$1, 2, 8, 9$ are all winning configurations, and top-3 has X_5 from which the majority of winning configurations go through. Top-$4, 5, 6, 7$ are losing configurations, something that can be validated by superimposing, for example, top-1 and top-4. The blue results also appear to be four rotations of a unique configuration, indicative of a potential common losing pattern. In a nutshell, $\hat{w}_{\bar{\bar{0}}}$ identifies interesting "red" and "blue" correlated sets that can act as latent factors for win and loss, respectively.

As a further experiment, we use estimators $\hat{w}, \hat{w}_0, \hat{w}_{\bar{0}}$ with exhaustive search. We report that $\hat{w}$ essentially orders the results according to cardinality, i.e., the top-1 is all the input variables $\mathcal{I}$, the next 9 are all subsets of $\mathcal{I}$ with size 9 etc. For $\hat{w}_0$ and $\hat{w}_{\bar{0}}$ there is agreement with the top 4 of $\hat{w}_{\bar{\bar{0}}}$, but the next 5 are all supersets of the top 2 with an extra cell. We find the results of $\hat{w}_{\bar{\bar{0}}}$ to be more interesting in this case.

VI. CONCLUSION

We considered the problem of measuring and efficiently discovering interpretable correlated sets from data. We adopted an information theoretic approach, and proposed a reliable and efficient estimator for normalized total correlation. In addition, we proposed effective algorithms for exhaustive and heuristic search, enabled by a tight bounding function.

For future work, we see many possibilities for extensions and improvements. Different estimators can be derived with appropriate algorithms. For example, the estimator of Vinh et al. [6] would allow incorporating prior knowledge to the problem. A conditional version of normalized total correlation would allow the discovery of correlated sets with respect to control variables. The algorithmic framework of Pennerath [17] for computing entropic measures could potentially be applied here to efficiently discover results for larger k values.

REFERENCES

[1] Y. Ke, J. Cheng, and W. Ng, "Correlated pattern mining in quantitative databases," *ACM Trans. Database Syst.*, vol. 33, pp. 14:1–14:45, 2008.

[2] X. Zhang, F. Pan, W. Wang, and A. Nobel, "Mining non-redundant high order correlations in binary data," *Proc. VLDB Endow.*, vol. 1, no. 1, pp. 1178–1188, Aug. 2008.

[3] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of Research and Development*, vol. 4, pp. 66 – 82, 02 1960.

[4] H.-V. Nguyen, P. Mandros, and J. Vreeken, "Universal dependency analysis," in *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016, pp. 792–800.

[5] S. Romano, N. X. Vinh, J. Bailey, and K. Verspoor, "A framework to adjust dependency measure estimates for chance," in *Proceedings of the 2016 SIAM international conference on data mining*, 2016, pp. 423–431.

[6] N. X. Vinh, J. Chan, and J. Bailey, "Reconsidering mutual information based feature selection: A statistical significance view," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2014.

[7] P. Mandros, M. Boley, and J. Vreeken, "Discovering reliable approximate functional dependencies," in *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 355–363.

[8] P. Mandros, M. Boley, and J. Vreeken, "Discovering Reliable Correlations in Categorical Data," *arXiv e-prints*, p. arXiv:1908.11682, Aug 2019.

[9] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[10] M. S. Roulston, "Estimating the errors on measured entropy and mutual information," *Physica D: Nonlinear Phenomena*, vol. 125, no. 3, pp. 285–294, 1999.

[11] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," vol. 11, no. Oct, pp. 2837–2854, 2010.

[12] H. Lancaster, *The chi-squared distribution*, ser. Probability and mathematical statistics. Wiley, 1969.

[13] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?" in *Proceedings of the 26th International Conference on International Conference on Machine Learning*. ACM, 2009, pp. 1073–1080.

[14] P. Mandros, M. Boley, and J. Vreeken, "Discovering reliable dependencies from data: Hardness and improved algorithms," in *IEEE International Conference on Data Mining*. IEEE, 2018.

[15] S. Romano, J. Bailey, N. X. Vinh, and K. Verspoor, "Standardized mutual information for clustering comparisons: One step further in adjustment for chance." 2014, pp. 1143–1151.

[16] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures & Algorithms*, vol. 19, no. 3-4, pp. 163–193, 2001.

[17] F. Pennerath, "An efficient algorithm for computing entropic measures of feature subsets," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2018, pp. 483–499.