Discovering robust dependencies from data

A dissertation submitted towards the degree

Doctor of Engineering (Dr.-Ing.)

of the Faculty of Mathematics and Computer Science

of Saarland University

by Panagiotis Mandros

Saarbrücken, 2020

Abstract

Science revolves around forming hypotheses, designing experiments, collecting data, and tests. It was not until recently, with the advent of modern hardware and data analytics, that science shifted towards a big-data-driven paradigm that led to an unprecedented success across various fields. What is perhaps the most astounding feature of this new era, is that interesting hypotheses can now be automatically discovered from observational data. This dissertation investigates knowledge discovery procedures that do exactly this. In particular, we seek algorithms that discover the most informative models able to compactly "describe" aspects of the phenomena under investigation, in both supervised and unsupervised settings.

We consider interpretable models in the form of subsets of the original variable set. We want the models to capture all possible interactions, e.g., linear, non-linear, between all types of variables, e.g., discrete, continuous, and lastly, we want their quality to be meaningfully assessed. For this, we employ information-theoretic measures, and particularly, the fraction of information for the supervised setting, and the normalized total correlation for the unsupervised. The former measures the uncertainty reduction of the target variable conditioned on a model, and the latter measures the information overlap of the variables included in a model.

Without access to the true underlying data generating process, we estimate the aforementioned measures from observational data. This process is prone to statistical errors, and in our case, the errors manifest as biases towards larger models. This can lead to situations where the results are utterly random, hindering therefore further analysis. We correct this behavior with notions from statistical learning theory. In particular, we propose regularized estimators that are unbiased under the hypothesis of independence, leading to robust estimation from limited data samples and arbitrary dimensionalities. Moreover, we do this for models consisting of both discrete and continuous variables.

Lastly, to discover the top scoring models, we derive effective optimization algorithms for exact, approximate, and heuristic search. These algorithms are powered by admissible, tight, and efficient-to-compute bounding functions for our proposed estimators that can be used to greatly prune the search space.

Overall, the products of this dissertation can successfully assist data analysts with data exploration, discovering powerful description models, or concluding that no satisfactory models exist, implying therefore new experiments and data are required for the phenomena under investigation. This statement is supported by Materials Science researchers who corroborated our discoveries.

Zusammenfassung

In der Wissenschaft geht es um Hypothesenbildung, Entwerfen von Experimenten, Sammeln von Daten und Tests. Jüngst hat sich die Wissenschaft, durch das Aufkommen moderner Hardware und Datenanalyse, zu einem Big-Data-basierten Paradigma hin entwickelt, das zu einem beispiellosen Erfolg in verschiedenen Bereichen geführt hat. Ein erstaunliches Merkmal dieser neuen Ära ist, dass interessante Hypothesen jetzt automatisch aus Beobachtungsdaten entdeckt werden können. In dieser Dissertation werden Verfahren zur Wissensentdeckung untersucht, die genau dies tun. Insbesondere suchen wir nach Algorithmen, die Modelle identifizieren, die in der Lage sind, Aspekte der untersuchten Phänomene sowohl in beaufsichtigten als auch in unbeaufsichtigten Szenarien kompakt zu "beschreiben".

Hierzu betrachten wir interpretierbare Modelle in Form von Untermengen der ursprünglichen Variablenmenge. Ziel ist es, dass diese Modelle alle möglichen Interaktionen erfassen (z.B. linear, nicht-lineare), zwischen allen Arten von Variablen unterscheiden (z.B. diskrete, kontinuierliche) und dass schlussendlich ihre Qualität sinnvoll bewertet wird. Dazu setzen wir informationstheoretische Maße ein, insbesondere den Informationsanteil für das überwachte und die normalisierte Gesamtkorrelation für das unüberwachte Szenario. Ersteres misst die Unsicherheitsreduktion der Zielvariablen, die durch ein Modell bedingt ist, und letztere misst die Informationsüberlappung der enthaltenen Variablen.

Ohne Kontrolle des Datengenerierungsprozesses werden die oben genannten Maße aus Beobachtungsdaten geschätzt. Dies ist anfällig für statistische Fehler, die zu Verzerrungen in größeren Modellen führen. So entstehen Situationen, wobei die Ergebnisse völlig zufällig sind und somit weitere Analysen stören. Wir korrigieren dieses Verhalten mit Methoden aus der statistischen Lerntheorie. Insbesondere schlagen wir regularisierte Schätzer vor, die unter der Hypothese der Unabhängigkeit nicht verzerrt sind und somit zu einer robusten Schätzung aus begrenzten Datenstichproben und willkürlichen-Dimensionalitäten führen. Darüber hinaus wenden wir dies für Modelle an, die sowohl aus diskreten als auch aus kontinuierlichen Variablen bestehen. Um die besten Modelle zu entdecken, leiten wir effektive Optimierungsalgorithmen mit verschiedenen Garantien ab. Diese Algorithmen basieren auf speziellen Begrenzungsfunktionen der vorgeschlagenen Schätzer und erlauben es den Suchraum stark einzuschränken. Insgesamt sind die Produkte dieser Arbeit sehr effektiv für die Wissensentdeckung. Letztere Aussage wurde von Materialwissenschaftlern bestätig.

Contents

1 Introduction **1**
1.1 Overview of contributions and outline 5

2 Information-theoretic dependency measures and estimation **9**
2.1 Information-theoretic measures 9
2.2 Estimation 13

3 Discovering robust functional dependencies **17**
3.1 Permutation mutual information and properties 27
3.2 Hardness of optimization 35
3.3 Admissible bounding functions for pruning 39
3.4 Optimization algorithms 44
3.5 Evaluation 49
3.6 Discussion and conclusions 66

4 Discovering robust totally correlated sets **71**
4.1 Normalized total correlation 74
4.2 Permutation normalized total correlation 76
4.3 Optimization algorithms 79
4.4 Evaluation 84
4.5 Discussion and conclusions 91

5 Functional Dependency Discovery from Mixed-Type Data **97**
5.1 Preliminaries 99
5.2 Consistent mixed mutual information estimation 101
5.3 Practical mixed data estimator 105
5.4 Robust functional dependency discovery from mixed data 111
5.5 Evaluation 113
5.6 Discussion and conclusions 120

6 Conclusion **127**
6.1 Discussion and future directions 128

Bibliography **131**

Glossary **149**

1 Introduction

The recent advances of modern hardware and data analytics in collecting, storing, and analyzing large volumes of data, have propelled science towards a paradigm shift with an unprecedented success: data-driven scientific discovery [HTT09]. Scientists, with the help of algorithms, can now analyze data to predict system states, test correlations between random variables from observations, learn unforeseen patterns in nature, and discover new scientific laws [MCGBK19]. Materials Science, for example, investigates interpretable models that identify and meaningfully describe the mechanisms behind physical and chemical processes such as the crystallization of composite semiconductors [GVL+15, GBV+17, SBG+20]. A compact representation of these mechanisms, e.g., in the form "these atomic material properties directly influence crystallization", can be used to improve accuracy and generalization of prediction algorithms, validate domain knowledge, and even lead to new findings, facilitating therefore advancements in material development such as more efficient photovoltaics and batteries.

Arriving at novel discoveries is a process that involves cycles of forming hypotheses, designing experiments, acquiring data, identifying models and assessing their quality. The cycles start with ideas and data exploration [Tuk80]. The goal of this dissertation is to assist practitioners in the knowledge discovery process. In partic-

ular, we propose algorithms that explore data and solve the following problems.

Problem 1 (**Unsupervised**). *Given observational data over the random variables $\mathcal{I} = \{X_1, \ldots, X_d\}$ of some process under investigation, discover the most informative models to "describe" that process.*

Problem 2 (**Supervised**). *Given observational data over the random variables $\mathcal{I} = \{X_1, \ldots, X_d\}$ and Y of some process, with Y being a target under investigation, discover the most informative models to "describe" Y.*

Despite being compactly stated, these are formidable data analysis tasks with challenging requirements in order to be effective. First, for the analysts to understand the results and reason about, the procedures need to be interpretable. For example, while deep neural networks pushed the boundaries in prediction to the point they can achieve better-than-human performance, they lack the ability to explain how and why. Second, the procedures must be able to capture all possible interactions, whether linear, non-linear, multivariate. Some high-quality models can be missed if, for example, only pairwise associations are being modeled. Third, the procedures should work with data of arbitrary types, e.g., discrete, continuous, as this is the case in many practical scenarios. Fourth, since we are only given observational data, principally answering questions about the underlying mechanisms behind some process requires robust inference techniques. Otherwise, spurious discoveries can lead to wrong conclusions. Lastly, the procedures should be able to efficiently discover the best models out of all possible models. This way analysts can be confident with moving the models to post-processing, or concluding that the cycle should restart with new experiments and data. Moreover, by accounting for multiple models, the analysts have the opportunity to investigate alternative descriptions. In summary, we seek interpretable exploratory procedures [Tuk77], free from parametric assumptions, statistically robust, exact, and efficient.

In this dissertation, we consider interpretable models in the form of subsets $\mathcal{X} \subseteq \mathcal{I}$ of the original variable set $\mathcal{I}$. In other words, we seek the parts of the process that are most informative and can be used as descriptions. Solutions to Problems 1 and 2 then consist of three parts: deriving scoring functions to assess the quality of candidate subsets $\mathcal{X} \subseteq \mathcal{I}$ while satisfying the first three requirements, using

inference techniques that satisfy the fourth, and designing efficient combinatorial optimization algorithms to discover the top scoring subsets $\mathcal{X}^* \subseteq \mathcal{I}$.

Regarding appropriate scoring functions, information-theoretic tools [Sha48] are already the perfect candidates. The Shannon entropy of a random variable quantifies the amount of uncertainty, or equivalently the amount of information, as the expected number of bits to transmit one symbol. Mutual information quantifies the amount of shared information between two random variables and can capture any type of relationship. Both can generalize to sets of random variables with arbitrary data types. With these two ingredients as building blocks, one can design scoring functions formalizing concepts such as statistical (in)dependence, relevancy, and redundancy, in a non-parametric way. It should not be a surprise that for certain applications, e.g., learning Chow–Liu trees [CL68] and feature selection [BPZL12], mutual information is the means to an optimal solution. Properties like these have popularized information theory in several scientific communities including neuroscience [TL18] and molecular biology [Ada04]. In this dissertation, we investigate information-theoretic scoring functions that can meaningfully assess model quality for Problems 1 and 2.

After obtaining such functions, the next step is to estimate them from observational data. That is, we derive these functions assuming the data generating process is known, but in practice we only have limited samples of that process. While inference from data is in general a challenging task, for our purposes it can go arbitrarily wrong. As an example, let us consider two discrete random variables X and Y that are statistically independent in the population, having true mutual information $I(X;Y) = 0$. To estimate I from data, we have to rely on estimators such as the plugin (maximum likelihood) $\hat{I}_{\text{pl}}$. Despite being consistent, i.e., given enough data $\hat{I}_{\text{pl}}$ can measure the true mutual information, with limited data it can easily be that $\hat{I}_{\text{pl}}(X;Y) > 0$, i.e., X and Y appear to be dependent. In fact, it is even possible that X and Y appear to be maximally dependent, i.e., one variable functionally determines the other. For our settings where we have to compare the estimates of all subsets $\mathcal{X} \subseteq \mathcal{I}$, this overestimation trivially leads to spurious discoveries and hence to wrong conclusions. The situation becomes more complicated when we consider variable sets of mixed types. While discretizing

continuous variables is a standard practice, it is not clear whether the discretized estimates are consistent with the population, nor what conditions are required for this. Hence, we need to derive appropriate estimators for our scoring functions.

Finally, after arriving at appropriate estimators, we need to efficiently solve the resulting combinatorial optimization problems for finding the best models with potentially huge search spaces (exponential in the size of $\mathcal{I}$). Not only exhaustive search is infeasible, but such problems are often NP-hard excluding polynomial time exact solutions. Hence, we need to design exhaustive optimization algorithms that are effective in practice. For situations where worst-case exponential complexity is not practical, we need approximate algorithms that come with result guarantees such that the data analysts can still draw conclusions.

Overall, we investigate in this dissertation the following research questions.

Question 1. *Given a process over random variables $\mathcal{I} = \{X_1, \ldots, X_d\}$ and (potentially) a target random variable Y, how can we meaningfully describe and compactly represent aspects of that process?*

To answer this question, we employ information theory and investigate scoring functions to assess the quality of candidate description models $\mathcal{X} \subseteq \mathcal{I}$. Information-theoretic measures are naturally interpretable, non-parametric, and in many cases theoretically justified. For candidate models $\mathcal{X} \subseteq \mathcal{I}$, we use the fraction of information $F(\mathcal{X}; Y)$ for the supervised scenario (Ch. 3), and the normalized total correlation $w(\mathcal{X})$ for the unsupervised (Ch. 4). Both capture all possible variable interactions and measure the quality in $[0, 1]$, with 0 and 1 implying statistical independence and maximum dependency, respectively.

Question 2. *Given observational data of a process, how can we robustly measure the true population value for the fraction of information and normalized total correlation?*

We answer this question by employing notions from statistical learning theory. In particular, we design estimators that are unbiased under the null hypothesis of independence by regularizing the plugin estimators. The regularizers are the biases under the null hypothesis, estimated non-parametrically as the expected

values of the plugin estimators across sample permutations. For the fraction of information we can efficiently perform all possible sample permutations, while for the normalized total correlation we rely on an upper-bound for efficiency. In addition, for sets of discrete and continuous variables, we derive a consistent mixed-data mutual information estimator based on partitioning techniques for Euclidean spaces (Ch. 5).

Question 3. *Given robust estimators for the fraction of information and normalized total correlation, how do we efficiently discover the top scoring models $\mathcal{X}^* \subseteq \mathcal{I}$ with guarantees?*

We answer this question with exact, approximate, and heuristic combinatorial optimization algorithms. To greatly reduce the search space, we derive pruning rules that are based on admissible, tight, and efficient-to-compute bounding functions for our proposed estimators. For the normalized total correlation, we show that admissibility is only enabled under a strict search space enumeration order. For exact search, we employ the branch-and-bound framework that is in practice very effective for hard problems. For cases where worst-case exponential complexity in not practical, the branch-and-bound framework can principally trade solution optimality for runtime efficiency in the form of approximation guarantees $\alpha \in (0, 1]$. Moreover, when solutions guarantees are not mandatory and/or very fast solutions are required, we employ the standard greedy algorithm for heuristic search, that in our case, produces near-optimal results.

1.1 Overview of contributions and outline

This dissertation presents knowledge discovery algorithms for Problems 1 and 2. Our goal is to assist data analysts with data exploration, discovering powerful description models, or concluding that no satisfactory models exist, implying therefore new experiments and data are required for the process under investigation. In Figure 1.1 we showcase our solution for Problem 2. Our algorithms can be found online[1] under MIT License, together with detailed descriptions on how to use them.

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

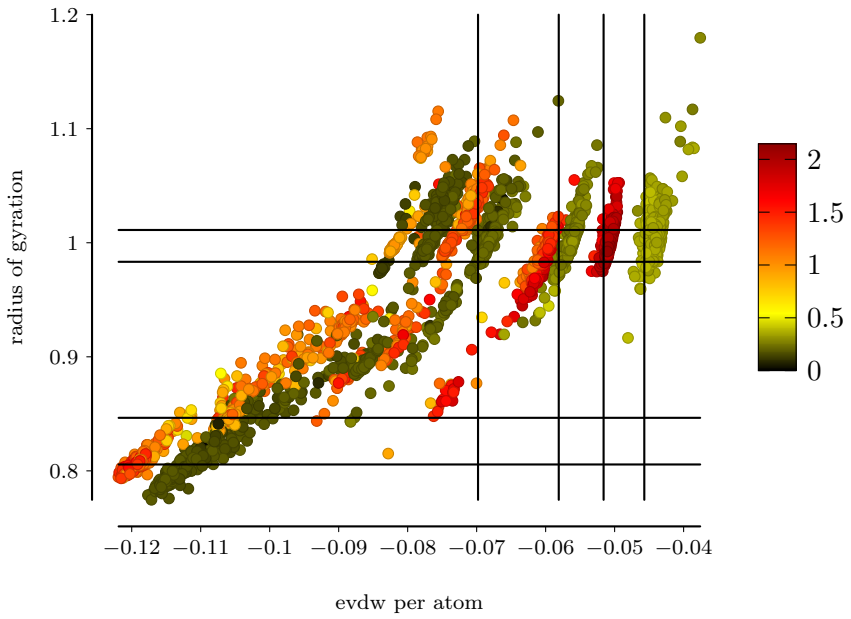


Figure 1.1: Demo of our solution to Problem 2 on a Materials Science case study. The dataset contains 12200 gold cluster configurations (of sizes 5 to 14 atoms) generated at finite temperature by replica-exchange molecular dynamics simulations [GBV+17]. The attributes in this dataset are 23 physicochemical and geometrical properties of the gold clusters. Here we are interested in discovering models that are descriptive for the target variable HOMO-LUMO gap that determines the electro-chemical properties of a cluster. Out of all possible $2^{22} - 1$ variable subsets, our proposed method uncovers as top $\mathcal{X}^*$ the structural variable "radius of gyration" and non-local dispersion energies "evdw per atom", that combined reduce the uncertainty of HOMO-LUMO gap by 43%, i.e., the estimated fraction of information is 0.43. The scatterplot represents the nano-clusters against the two-dimensional descriptor, with color indicating the values of the HOMO-LUMO gap. The lines represent the partition of $\mathbb{R}^2$ in up to 5 bins per axis that contributes the most to reducing the uncertainty of HOMO-LUMO gap.

Below is an overview of the contributions that appear in Chapters 3, 4, and 5. The dissertation starts with an introduction to information-theoretic measures and estimation in Chapter 2, and finishes with discussion and conclusions in Chapter 6.

Supervised knowledge discovery. We conclude in Chapter 3 that maximizing the fraction of information $F(\mathcal{X}; Y)$ over all subsets $\mathcal{X} \subseteq \mathcal{I}$ is the desired solution for Problem 2. The fraction of information quantifies the proportional

reduction of uncertainty of random variable Y by knowing $\mathcal{X}$. It takes values in $[0, 1]$, with extreme values indicating statistical independence and functional dependency $Y = f(\mathcal{X})$, respectively. Besides interpretability, this is a theoretically justified goal. The top $\mathcal{X}^* \subseteq \mathcal{I}$ achieves the Bayes error in classification, it is the solution to maximizing the conditional likelihood, and if we assume a Bayesian network, it has causal interpretations. We then derive an estimator for the fraction of information and discrete data that is corrected for inflated estimates and well-suited for high-dimensional optimization. This estimator is the difference of the plugin estimator versus its average value across all possible sample permutations, and is essentially unbiased under the hypothesis of independence. We show that the problem of maximizing this estimator is NP-hard, justifying both exhaustive and heuristic search. We derive bounding functions for the estimator that can be used as pruning rules in both styles of optimization. Last, we propose algorithms for exact, approximate, and heuristic search. The results demonstrate that the resulting method is efficient, statistically robust, and that it indeed discovers meaningful dependencies. This chapter also lays the foundation for the following chapters, e.g., by introducing algorithms, pruning rules, and estimators. This chapter is based on work published as Mandros et al. [MBV17, MBV18, MBV20].

Unsupervised knowledge discovery. We argue in Chapter 4 that maximizing the normalized total correlation $w(\mathcal{X})$ among all $\mathcal{X} \subseteq \mathcal{I}$ is an attractive solution for Problem 1. Normalized total correlation takes values in $[0, 1]$, quantifying the proportion of mutual dependency residing in a set of random variables compared to the maximum possible. It takes value 0 for statistical independence between all $X \in \mathcal{X}$, and value 1 if there exists a $X \in \mathcal{X}$ such that $X' = f(X)$ for all the remaining $X' \in \mathcal{X} \setminus \{X\}$. We then derive a fast-to-compute estimator for the normalized total correlation and discrete data that is corrected for inflated estimates and well-suited for high-dimensional optimization. This estimator is based on subtracting an upper-bound for the expected value across all sample permutations. We derive bounding functions for the estimator to be used as pruning rules in combinatorial optimization. For the bounding functions to be admissible, we identify a strict enumeration order. Last, we propose algorithms for exact, approximate, and heuristic search. The results demonstrate that the resulting

method is efficient, statistically robust, and that it indeed discovers meaningful dependencies. This chapter is based on work published as Mandros et al. [MBV19].

Estimating mutual information from mixed-type data. It is known that the population mutual information $I(X;Y)$ between a pair of continuous random variables X and Y can be attained non-parametrically as the limit of a series of finer-grained equal-width quantizations. That is, if we quantize the domain of the continuous random variables X and Y in $k \in \mathbb{Z}^+$ and $q \in \mathbb{Z}^+$ equal-width bins to create discrete variables X_k, Y_q, respectively, then it holds that $I(X;Y) = \lim_{k,q\to\infty} I(X_k, Y_q)$. We extend this result for sets of variables $\mathcal{X}$ and $\mathcal{Y}$ that can be mixtures of random variables of any type, as well as identify a larger class of quantization techniques applicable. We then show how to apply this process for functional dependency discovery given observational data, arriving at a mixed-data mutual information estimator framework that requires two ingredients: a discrete consistent estimator and a partitioning technique. We argue that not all consistent estimators can achieve robust estimation. Last, we show that this estimator framework synergizes well with our proposed method for robust functional dependency discovery (Ch. 3). In particular, the robust estimator allows polynomial time exact search through all possible Euclidean space partitions, and efficient pruning rules can be derived. Results demonstrate that the estimation process is indeed robust and consistent, while the resulting discovery algorithms remain effective for data of arbitrary variable types. These appear in Chapter 5, which is based on work published as Mandros et al. [MKBV20].

2

Information-theoretic dependency measures and estimation

In this chapter we cover the information-theoretic tools we use throughout the dissertation. We show how to estimate these tools given empirical data, and we introduce basic notation.

2.1 Information-theoretic measures

Ever since Claude E. Shannon formalized information theory in his 1948 seminal article "A mathematical theory of communication" [Sha48], information-theoretic principles have been adopted virtually in every machine learning and data mining application. Examples include decision trees [Qui86], feature selection [Lew92], representation learning [Lin88, TPB00], learning tree-structured distributions [CL68], computer vision [VWI97]. The reason for being so broadly applicable is that information theory provides tools that meaningfully quantify the uncertainty of random variables, or equivalently, the amount of information.

The **Shannon entropy** for a discrete random variable Y, with domain V_Y,

domain size $S_Y = |V_Y|$, and probability distribution p_Y, is defined as the functional

$$H(Y) = -\sum_{y \in V_Y} p_Y(y) \log(p_Y(y)) \ .$$

Note that we use $y \in Y$ and $p(y)$ instead of $y \in V_Y$ and $p_Y(y)$, respectively, whenever clear from the context. Assuming a logarithm of base 2, Shannon entropy quantifies uncertainty as the expected number in bits to transmit one symbol from Y. For example, assuming that Y follows a uniform probability, then it has the maximum possible[1] entropy of $\log(S_Y)$. Given a second random variable X and probability distribution $p_{X,Y}$, the **conditional Shannon entropy** of Y given X is defined as

$$\begin{aligned} H(Y \,|\, X) &= \sum_{x \in X} p(x) H(Y \,|\, X = x) \\ &= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y \,|\, x) \log(p(y \,|\, x)) \ . \end{aligned}$$

The amount of shared information between the two variables is quantified by the **mutual information** functional

$$\begin{aligned} I(X;Y) =& \sum_{x \in X} \sum_{Y \in Y} p(x,y) \frac{p(x,y)}{p(x)p(y)} \\ =& H(Y) - H(Y \,|\, X) \\ =& H(X) - H(X \,|\, Y) \\ =& H(X) + H(Y) - H(X,Y) \ . \end{aligned}$$

Mutual information satisfies an important property useful in many data analysis tasks: $I(X;Y) \geq 0$, with equality if and only if X and Y are statistically independent, a fundamental relation we recall here for self-containment.

Definition 2.1.1 (**Statistical independence**)**.** *For two random variables X and Y, we say that X and Y are statistically independent, denoted as $X \perp\!\!\!\perp Y$, if and*

[1] Among all distributions with the same domain size.

only if for every $x \in X$ and $y \in Y$, it holds that $p(x, y) = p(x)p(y)$.

Note that mutual information is the Kullback–Leibler divergence [KL51] between $p(X, Y)$ and $p(X)p(Y)$. Now given a third variable Z, the **conditional mutual information** of X and Y given Z is defined as

$$\begin{aligned} I(X;Y \mid Z) =& H(Y \mid Z) - H(Y \mid Z, X) \\ =& H(X \mid Z) - H(X \mid Z, Y) \ . \end{aligned}$$

The last measure we introduce is the **fraction of information**, an asymmetric normalized version of mutual information expressed as

$$F(X;Y) = \frac{I(X;Y)}{H(Y)} \ .$$

The fraction of information is a supervised measure quantifying the proportional reduction of uncertainty of a target variable (the second argument) by conditioning on another. It takes values in $[0, 1]$, with extreme values indicating statistical independence and **functional dependency**, respectively. Here, functional dependency is defined as follows.

Definition 2.1.2 (**Functional dependency**)**.** *For two random variables X and Y, we say that Y functionally depends on X, denoted as $Y = f(X)$, if and only if for every value $y \in Y$, there exists a value $x \in X$ such that $p(y \mid x) = 1$.*

The following proposition summarizes important properties satisfied by the aforementioned information-theoretic measures. In Figure 2.1 we illustrate their relationships with Venn diagrams.

Proposition 2.1.1 ([CT06], Ch. 2)**.** *Given random variables X, Y, Z, the following statements hold:*

a) $H(Y) \geq H(Y \mid X)$ with equality if and only if $Y \perp\!\!\!\perp X$

b) $H(Y \mid X) = 0$ if and only if Y is a function of X

c) $I(X;Y) \geq 0$ with equality if and only if $X \perp\!\!\!\perp Y$

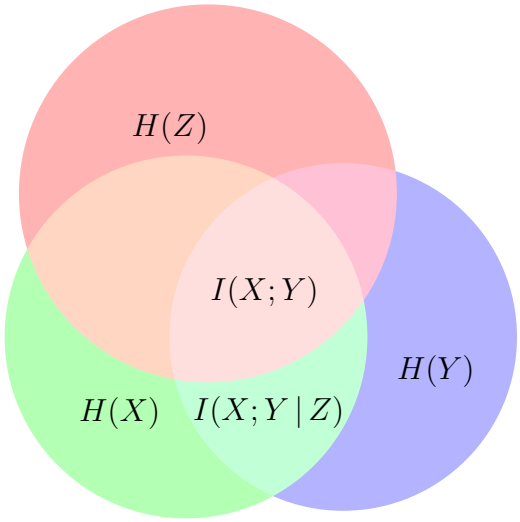


Figure 2.1: Venn diagram for relations among information-theoretic measures. The red, green, and blue circles are the Shannon entropies of Z, X, Y, respectively, while the rugby ball shaped region in the middle is the mutual information of X and Y, and the small light green is $I(X;Y \mid Z)$. The purple area is $H(Y \mid X, Z)$. All the circles combined is $H(Z, X, Y)$. No overlap would imply $H(Z, X, Y) = H(Z) + H(X) + H(Y)$, and 0 mutual information for all pairs.

d) $I(X;Y) \leq \min\{H(X), H(Y)\}$

e) $I(X;Y \mid Z) = 0$ *if and only if* $X \perp\!\!\!\perp Y \mid Z$

f) $F(X;Y) \in [0,1]$, *with* 0 *if and only if* $X \perp\!\!\!\perp Y$, *and* 1 *if and only if* $Y = f(X)$

In addition to these, we note the following desired properties:

- these measures, as functionals of probability distributions, trivially extend to sets of variables and multivariate distributions. For example, given sets of variables $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_l\}$, with values $\mathbf{x} \in V_{\mathcal{X}}, \mathbf{y} \in V_{\mathcal{Y}}$, and probability distribution $p_{\mathcal{X},\mathcal{Y}}$, we have $I(\mathcal{X};\mathcal{Y})$, $H(\mathcal{X}), H(\mathcal{X},\mathcal{Y})$, etc.
- Moreover, while so far we considered the discrete case, these measures are agnostic with respect to variable types. That is, the random variables involved can be nominal, ordinal, continuous, etc. We explore this further in Chapter 5.

- Mutual information, and hence the fraction of information, is agnostic with respect to the variable relationship, e.g., non-linear, XOR. This is easy to verify since mutual information detects dependence as lack of independence, without the need to assume specific relationship types.

To summarize, information theory provides a high-level language to build objective functions and formalize concepts such as statistical (in)dependence, relevancy, redundancy, in a non-parametric way. Note, however, that this language involves probability distributions which we assumed so far to be known. That is, we were at **population level**. In practice, we estimate them from the **empirical level** with data-dependent estimators $\hat{H}, \hat{I}, \hat{F}$.

2.2 Estimation

We consider in this dissertation two scenarios: an unsupervised, where we are given n i.i.d. samples $\mathbf{D}_n = \{\mathbf{d}_1, \ldots, \mathbf{d}_n\}$ over input random attributes $\mathcal{I} = \{X_1, \ldots, X_d\}$ with joint probability distribution $p(\mathcal{I})$, and a supervised, with an additional special random variable of interest Y and joint probability distribution $p(\mathcal{I}, Y)$. In both cases, the goal is to maximize some information-theoretic measure over all possible subsets $\mathcal{X} \subseteq \mathcal{I}$. We focus on discrete random variables, with the mixed-type scenario explored in Chapter 5.

We identify samples of a random variable X with the **labeling** $X\colon [n] \to V_X$ it induces on the data sample, i.e., $X(i) = \mathbf{d}_i(X)$. Moreover, for a set $\mathcal{X} = \{X_1, \ldots, X_l\}$ of labelings over $[n]$, we define the corresponding vector-valued labeling by $\mathcal{X}(i) = (X_1(i), \ldots, X_l(i))$. We define $c_X : V_X \to \mathbb{Z}^+$ to be the **empirical counts** of X, i.e., $c_X(x) = |\{i \in [n]\colon X(i) = x\}|$. We further denote with $\hat{p}_X : V_X \to [0,1]$, where $\hat{p}_X(x) = {}^{c_X(x)}/_n$, the **empirical distribution** of X. Given another random variable X', $\hat{p}_{X \mid X'=x'} : V_X \to [0,1]$ is the **empirical conditional distribution** of X given $X' = x'$, with $\hat{p}_{X|X'=x'}(x) = {}^{c_{X,X'}(x,x')}/_{c_{X'}(x')}$ for $x \in X$. However, we use $\hat{p}(x)$ and $\hat{p}(x \,|\, x')$, respectively, whenever clear from the context.

Straightforward estimators for H, I, F, can be derived by maximum likelihood estimation, where the empirical distribution $\hat{p}$ is plugged in to evaluate the function-

als. This gives rise to **plugin** estimators $\hat{H}_{\text{pl}}$, $\hat{I}_{\text{pl}}$, and $\hat{F}_{\text{pl}}$. Despite their simplicity, these estimators are **consistent** [AK01], i.e., they converge in probability to the true population values as $n \to \infty$. While this is a desired statistical property, the research community has argued that these estimators are of little practical use in the case of high-dimensional distributions, or equivalently, large alphabet distributions.[2] This statement is nicely captured with the following quote[3] by Wyner and Foster [WF03]:

> "the plugin estimate is universal and optimal not only for finite alphabet i.i.d. sources but also for finite alphabet, finite memory sources. On the other hand, practically as well as theoretically, these problems are of little interest."

The importance of information theory and the need for better estimators has therefore led to a large amount of proposals [NSB02, Gra08, SG96, Gra88, VV13, WY16, VV11, JVHW15, Pan03, Pan04, Sch13] (see Jiao et al. [JVHW15] for a great review). Here we focus on two theoretically optimal estimators, namely the **unseen estimator** $\hat{H}_{\text{un}}$ by Valiant and Valiant [VV13] and the **minimax estimator** $\hat{H}_{\text{mm}}$ by Jiao et al. [JVHW15]. The former tries to estimate the unseen portion of the population[4] solving linear programs, while the second employs best-polynomial approximation to obtain a minimax rate-optimal estimator. These two are theoretically optimal because they achieve the necessary and sufficient conditions for consistent estimation of entropy, i.e., they require number of samples that are sublinear to the domain size of the variable involved. In more detail, the **sample complexity** of an estimator represents the minimum number of samples required to achieve a certain ϵ-δ-PAC guarantee. This is a more useful evaluation criterion than limit consistency. For Shannon entropy, the sample complexity is expressed as a function of the domain size of the random variable involved, with the plugin estimator $\hat{H}_{\text{pl}}$ requiring number of samples linear to the domain size,

[2] A set of random variables $\mathcal{X} = \{X_1, \ldots, X_m\}$ can be seen as a single variable X with domain $V_X = \prod_{X_i \in \mathcal{X}} V_{X_i}$. In that sense, large alphabet and high-dimensional are equivalent.

[3] Which came into our attention by being quoted in Jiao et al. [JVHW15].

[4] Essentially seek a better estimate for p rather than simply using the empirical $\hat{p}$.

i.e., $S_{\hat{H}_{\text{pl}}}(k) \in O(k)$, where k is the domain size. The unseen and the minimax achieve the best possible $S_{\hat{H}}(k) \in \Omega(k/\log(k))$.

The majority of the estimators proposed, including the theoretically optimal, have not been used in statistically high-demanding practical scenarios such as the one we consider in this dissertation, i.e., finding the maximum over all subsets $\mathcal{X} \subseteq \mathcal{I}$. In the following chapters we show that they indeed under-perform, and proceed to derive appropriate estimators.

3

Discovering robust functional dependencies

Given categorical data over attributes $\mathcal{I} = \{X_1, \ldots, X_d\}$ and a target attribute of interest Y, there exist a multitude of applications concerned with discovering subsets $\mathcal{X} \subseteq \mathcal{I}$ that are "relevant" to Y. Hence, it is a central research topic in many communities including data management, feature selection, Bayesian networks, and causal inference. To arrive at a desired solution for our knowledge discovery purposes, let us first review the different methods proposed.

In **data management**, practitioners seek (approximate) functional dependencies [RG99, Ch. 15], also known as keys, to be used in applications such as schema discovery [KPHN16], data integration [MHH+01], schema design [KLLZ16], normalization [PN17], query relaxation [NK04]. Here, a functional dependency is a mapping from the values of some $\mathcal{X} \subseteq \mathcal{I}$ to Y, i.e., the values of $\mathcal{X}$ uniquely determine the values of Y, while an approximate functional dependency allows for errors. A plethora of methods have been proposed (see, e.g., [PEM+15, LLLC12]),

This chapter is an extended version of work that originally appeared in ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) [MBV17], IEEE International Conference on Data Mining (ICDM) [MBV18], and Knowledge and Information Systems (KAIS) [MBV20].

that in a nutshell, quantify the degree of functional dependency by counting the amount of data samples violating it and search for all minimal subsets satisfying some threshold. The fraction of information has also been proposed to better quantify the degree of functional dependency [CP87, GR04]. Regarding knowledge discovery, this type of analysis has a major drawback due to the implicit closed-world assumption where the data generating process p is equal to the empirical $\hat{p}$ [GR04]. Hence, the dependencies discovered, although functional, reflect only the structure of the given empirical data and do not generalize.

Feature selection is concerned with selecting a subset of attributes to facilitate prediction algorithms in terms of accuracy, generalization, training/testing time, as well as reduce storage requirements [GE03, LY05, JKP94, LMSZ10]. Hence, they must be able to efficiently provide subsets of attributes that contain the majority of predictive information for Y. The various methods proposed can be split into three categories: wrapper, embedded, and filter methods. Wrappers select features by using an induction algorithm as a black box, and via various search strategies, find features that maximize some performance metric, e.g., cross-validation error [KJ97]. Embedded methods are prediction algorithms with an embedded feature selection process, e.g., LASSO with the $L1$ penalty term for sparsity [Tib96]. **Filter methods**, in contrast to the previous two, select features independently of any prediction algorithm by using proxy measures to assess the quality of feature subsets [KR92, Lew92]. The most prominent feature selection methods are filters based on information-theoretic scoring functions maximized with cardinality constraints (desired number of features) by the greedy algorithm [Lew92, Bat94, VE14, BPZL12]. This comes as no surprise since filter methods are independent of any prediction algorithm, mutual information captures arbitrary relationships and is invariant under invertible and differentiable transformations, while the greedy algorithm is efficient. Some well-known methods are presented in Table 3.1. Given the set of already selected features $\mathcal{S}$, these algorithms select the next best feature X^* by maximizing a univariate target relevance term, minus the redundancy with $\mathcal{S}$ modeled as a function of low-order (e.g., pairwise for MRMR) interactions. It is worth noting that Brown et al. [BPZL12] unify in seminal work two decades of research, showing that pro-

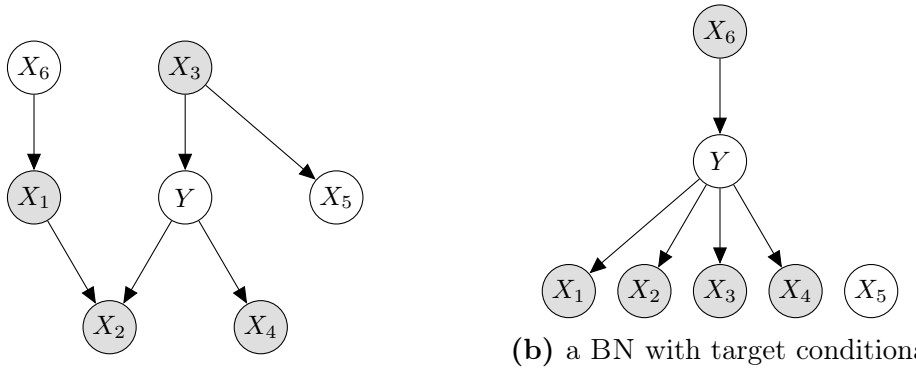
Table 3.1: Representative sample of filter feature selection algorithms based on information theory. With $\mathcal{I} = \{X_1, \ldots, X_d\}$ we denote the set of features, Y is the target variable, and $\mathcal{S}$ is the set of already selected features, i.e., if we assume we currently seek the k-th feature, $\mathcal{S}$ contains the previous $k-1$ selected features. See [BPZL12, VE14] for extensive reviews.

method	objective function
MRMR [PLD05]	$\arg\max_{X \in \mathcal{I} \setminus \mathcal{S}} \left(I(X;Y) - \frac{1}{\lvert\mathcal{S}\rvert} \sum_{X' \in \mathcal{S}} I(X;X') \right)$
CMIM [Fle04]	$\arg\max_{X \in \mathcal{I} \setminus \mathcal{S}} \Big(I(X;Y) - \max_{X' \in \mathcal{S}} \left(I(X;Y) - I(X;Y \mid X') \right) \Big)$
CIFE [LT06]	$\arg\max_{X \in \mathcal{I} \setminus \mathcal{S}} \Big(I(X;Y) - \sum_{X' \in \mathcal{S}} \left(I(X;Y) - I(X;Y \mid X') \right) \Big)$
JMI [YM99]	$\arg\max_{X \in \mathcal{I} \setminus \mathcal{S}} \Big(I(X;Y) - \frac{1}{\lvert\mathcal{S}\rvert} \sum_{X' \in \mathcal{S}} \left(I(X;Y) - I(X;Y \mid X') \right) \Big)$

posed information-theoretic filter methods are approximations to the problem of maximizing mutual information under certain assumptions.[1] In fact, maximizing mutual information (and not low-order approximations) is considered to be the ultimate feature selection procedure with theoretical justifications: it leads to the Bayes error in classification [FH61, Ch. 9] [HR70, TD68], i.e., the minimum possible classifier-independent error, and it corresponds to maximizing the conditional likelihood of the target given subsets of attributes [BPZL12]—a fundamental principle in statistics. Despite their success in feature selection, filter methods are not tailored towards knowledge discovery: they do not consider functional dependencies and hence potential high-order relationships can be missed, while the necessary assumptions for both statistical and computational efficiency restrict the space of possible hypotheses for the data generating distribution (see Fig. 3.1).

Markov blanket discovery algorithms is a family of methods that identify local neighborhoods in Bayesian networks. A **Bayesian network (BN)** [Pea88, Ch. 3.2] is a directed acyclic graph (DAG) over a set of variables corresponding to a factorization of a joint probability distribution and for which the Markov

[1]For example, an assumption for MRMR is that of target conditional independence, the same assumption as that of Naive Bayes (see Figure 3.1b).



(a) a general BN

(b) a BN with target conditional independence assumption

Figure 3.1: Examples of two Bayesian networks over variables $\mathcal{I} = \{X_1, \ldots, X_6\}$ and target Y. Shaded nodes indicate the Markov blanket of Y. **Left**: a general BN where $\mathrm{MB}(Y)$ comprises of the parents of Y (i.e., X_3), the children (i.e., X_2, X_4), and the parents of the children (i.e., X_1, also known as spouses). **Right**: a BN that has the structural assumption of target conditional independence. Note that with this assumption Y can have at most 1 parent.

condition[2] holds: every variable Z in the DAG is conditionally independent of its non-descendants[3] given its parents.[4] The Markov blanket of a random variable is defined as follows.

Definition 3.0.1 (**Markov blanket (MB)**). *In a Bayesian network over set of random variables $\mathcal{R}$, the Markov blanket of a node $Y \in \mathcal{R}$, which we denote as $\mathrm{MB}(Y)$, is a minimal subset of features $\mathcal{S} \subseteq \mathcal{R} \setminus \{Y\}$ for which it holds that $Y \perp\!\!\!\perp Z \,|\, \mathcal{S}$, for all $Z \in \mathcal{R} \setminus (\mathcal{S} \cup \{Y\})$. In other words, $\mathrm{MB}(Y)$ is a set of variables that renders Y independent of the remaining $\mathcal{R}$. See Figure 3.1 for examples.*

As the definition suggests, Markov blankets are (at least from a BN point of view) the optimal set of attributes. Interestingly, Markov blankets can be easily identified under the assumption of **faithfulness**: the only independencies to hold in the distribution are those entailed by the Markov condition,[5] and as

[2]Also known as local Markov condition and Markov assumption.

[3]Nodes that cannot be reached by directed paths starting from Z.

[4]Nodes that have a directed edge towards Z.

[5]An example of a faithfulness violation is the XOR relationship $Y = X \oplus Z$ for uniform binary X and Z, where the DAG is $X \to Y \leftarrow Z$ with $X \not\perp\!\!\!\perp Z$ as the only independence, while we

a consequence, it guarantees the existence of a unique MB for every variable comprising of the parents, children, and spouses. Note that Markov blankets are devoid of directionality, i.e., knowing the MB does not tell us which nodes are the parents, etc. Established approaches that operate under this assumption, e.g., IAMB [TA03, TASS03], GS [MT99], or the more data efficient[6] HITON [ATS03], MMMB [TAS03], PCMB [PNBT07], can be abstracted as grow-shrink type of algorithms that employ conditional independence tests to first identify the true positives (i.e., the members of the MB), and then remove the false positives. In Algorithm 1 we present IAMB that implements the grow step with the while loop and the shrink step with the for loop. If the faithfulness assumption is violated, and particularly the intersection property,[7] then there can exist multiple Markov blankets. For this, various randomized and approximate methods have been proposed for weaker assumptions,[8] e.g., KIAMB [PNBT07], EGSG [LLZ10], TIE [SLA13]. It is important to note that Markov blanket discovery algorithms, unlike filters, explicitly consider a data generating model (i.e., the BN) which they can recover (without directionalities) under certain non-structural assumptions (e.g., faithfulness), and hence they can give insights about potential underlying mechanisms.[9] However, they exhibit the following drawbacks. First, the faithfulness assumption implies that not all possible relationships are considered, e.g., XOR. Second, to retrieve multiple Markov blankets they resort in randomized and approximate techniques. Third, the results heavily depend on parameters such as significance level for the tests, as well as a parameter limiting the size of the conditional set for efficiency and statistical robustness.

From a **causal inference** perspective, identifying relevant attributes for Y in data corresponds to identifying causal relationships: we say that X causes Y if an intervention on the underlying data generating process to assign different values

additionally have in the distribution that $X \perp\!\!\!\perp Y$ and $Z \perp\!\!\!\perp Y$.

[6]They are more data efficient because they employ conditional independence tests with subsets of the currently selected Markov blanket, while IAMB and GS condition on the entire MB.

[7]That is, the distribution is not strictly positive.

[8]For example, the composition property.

[9]There exists work relating feature selection and Markov blankets, e.g., the MB comprises of the strongly relevant features [JKP94] in feature selection for faithful distributions [PE08, GAE07, AST+10].

Algorithm 1 IAMB: Given a set of input variables $\mathcal{I}$, target Y, independence test T, significance level α, and dependency score Q, the algorithm returns under the faithfulness assumption the Markov blanket of Y

function IAMB($\mathcal{I}, Y$)
  MB $= \emptyset$
  while MB does not change **do**
    $X^* = \arg\max\{Q(X; Y) \colon X \in \mathcal{I} \setminus \text{MB}\}$
    if $X^* \not\perp\!\!\!\perp Y \mid \text{MB}$ **then** ▷ according to T and α
      MB $=$ MB $\cup \{X^*\}$
    $\mathcal{I} = \mathcal{I} \setminus \{X^*\}$
  for $X \in$ MB **do**
    if $X \perp\!\!\!\perp Y \mid \text{MB} \setminus \{X\}$ **then** ▷ according to T and α
      MB $=$ MB $\setminus \{X\}$
  return MB

x to X, causes the conditional probabilities $p(Y \mid do(X = x))$ to change. These are not associative relationships like the aforementioned scenarios, but rather reflect the underlying mechanisms of the system under investigation. While the former can be solved solely by employing classic statistics on i.i.d. samples, e.g., an independence test, the gold standard to identify causal relationships are carefully designed randomized control trial experiments that are, however, in many cases prohibitive. For example, we cannot force people to adopt bad habits in order to test whether they cause bad consequences. Hence, in such situations, causality must be inferred from observational data and Bayesian networks are augmented with necessary assumptions for identifiability[10] (see, e.g., [Pea09, Ch. 2] [Pea88, Ch. 8]). For example, to give a causal interpretation to Markov blankets, besides faithfulness, there are two additional assumptions required. One assumption is the belief that the DAG represents causal relationships, which translates the Markov condition to the causal Markov condition, i.e., every variable is conditionally independent of its non-descendants given its direct causes.[11] A second assumption

[10] Intuitively, the goal is ensure that the directionality in the DAG matches the true causal structure, and with more assumptions, the stronger the claims can be.

[11] In a causal DAG, the notions of parents, children, spouses, correspond to direct causes, direct effects, and direct causes of the direct effects, respectively.

is that of causal sufficiency which states that all common causes for any pair of variables are measured. Under these three assumptions, the (unique) Markov blanket is the set of direct causes, direct effects, and direct causes of the direct effects. Note that by employing conditional independence tests, one gets an equivalence class of oriented graphs[12] (Markov equivalent graphs). For stronger claims about directionality, the class of variable relationships is restricted, e.g., using additive noise models [HJM+09, PMJS14].

To meet our goal for an effective knowledge discovery, we investigate in this chapter a different angle. In particular, we consider the combinatorial optimization problem of maximizing the fraction of information for the top-k, i.e., find the k attribute subsets $X_1^*, \ldots, X_k^* \subseteq \mathcal{I}$ that satisfy

$$F(X_i^*; Y) = \max\{F(\mathcal{X}; Y) \colon\ F(\mathcal{X}_{i-1}^*; Y) \geq F(\mathcal{X}; Y), \mathcal{X} \subseteq \mathcal{I}\}\ . \tag{3.1}$$

Note that since $H(Y)$ is constant during the optimization, this problem is equivalent to maximizing mutual information. We call Eq.(3.1) the **functional dependency discovery (FDD)** task,[13] which is well-motivated for our goal. First of all, it enjoys all the benefits of using the fraction of information: it is agnostic of the relationship type, agnostic of the type of variables involved (e.g., discrete, continuous), and is interpretable as it quantifies in $[0, 1]$ the relative reduction of uncertainty of Y by knowing $\mathcal{X}$, with extreme values indicating statistical independence and functional dependency, respectively. Second, it is theoretically justified as it satisfies optimality criteria such as Bayes error and conditional likelihood maximization equivalence. Third, while it is free from structural assumptions regarding the data generating process p (e.g., a DAG), by making the assumption that p is a BN solutions to Eq.(3.1) can correspond to Markov blankets without needing additional assumptions such as faithfulness.

[12] There are more than one directed graphs matching the independencies found in the distribution with different directionalities.

[13] Note that this term is standard terminology in data management, and we are in fact interested in the same goal, i.e., finding approximate keys. In our case, however, we aim to recover keys w.r.t. the true data generating process p.

Theorem 3.0.1. *Given random variables $\mathcal{I} = \{X_1, \ldots, X_d\}$ and Y with joint probability distribution represented by a Bayesian network where Y has k Markov blankets, the top-k minimal and maximal solutions to Eq.*(3.1) *correspond to the k Markov blankets of Y and vice versa.*

Proof. Here, minimality ensures that the top-k does not include sets $\mathcal{X}, \mathcal{X}'$, with $\mathcal{X} \subseteq \mathcal{X}'$ and $F(\mathcal{X}; Y) = F(\mathcal{X}'; Y)$. In that case, only $\mathcal{X}$ is part of the top-k. This can be achieved, e.g., by breadth-first search. Maximality ensures that the top-k does not not include sets $\mathcal{X}, \mathcal{X}'$, with $\mathcal{X} \subseteq \mathcal{X}'$ and $F(\mathcal{X}; Y) < F(\mathcal{X}'; Y)$. In that case only $\mathcal{X}'$ is part of the top-k. This could be achieved by growing a prefix tree for the top-k, where only the root-to-leaf paths are reported. Then, the proof follows directly from the definition of a Markov blanket and the properties of mutual information. We prove both directions by contradiction.

Let us assume that at least one of the maximizers $\mathcal{X}_i^*, i \in [k]$, is not a Markov blanket. Then there exists a $Z_i \in \mathcal{I} \setminus \mathcal{X}_i^*$ for which $Y \not\perp\!\!\!\perp Z_i \,|\, \mathcal{X}_i^*$. We know that mutual information is monotonically increasing with the superset relation, i.e., $I(\mathcal{X}; Y) \leq I(\mathcal{X}'; Y)$ for $\mathcal{X} \subseteq \mathcal{X}' \subseteq \mathcal{I}$, with equality if only if $Y \perp\!\!\!\perp (\mathcal{X}' \setminus \mathcal{X}) \,|\, \mathcal{X}$. Therefore, adding the Z_i in $\mathcal{X}_i^*$ would result in $I(\mathcal{X}_i^* \cup \{Z_i\}; Y) > I(\mathcal{X}_i^*; Y)$, which is a contradiction as then $\mathcal{X}_i^*$ would not be part of the top-k. Conversely, let us assume that there exists at least one Markov blanket that is not a top-k maximizer. As every maximizer is a Markov blanket, this would imply there exist at least $k+1$ Markov blankets, which is a contradiction as there are k Markov blankets. □

The result implies that Eq.(3.1) can be seen as a score-based approach[14] for learning Markov blankets in Bayesian networks, and hence the solutions have structural and causal interpretations. In fact, Eq.(3.1) is potentially a superior formulation, at least in theory, as it does not require further assumptions and can retrieve and assess the quality of multiple Markov blankets.

Despite the theoretical justifications, solving Eq.(3.1) is a very challenging problem in practice. Estimating the fraction of information from an empirical

[14]Note that score-based in Bayesian structure learning terminology corresponds to methods that infer the Markov equivalent DAG, and not only Markov blankets, by scoring the different DAG with functions such as maximum likelihood and the Bayesian Information Criterion [GH94, HMC06, Chi03].

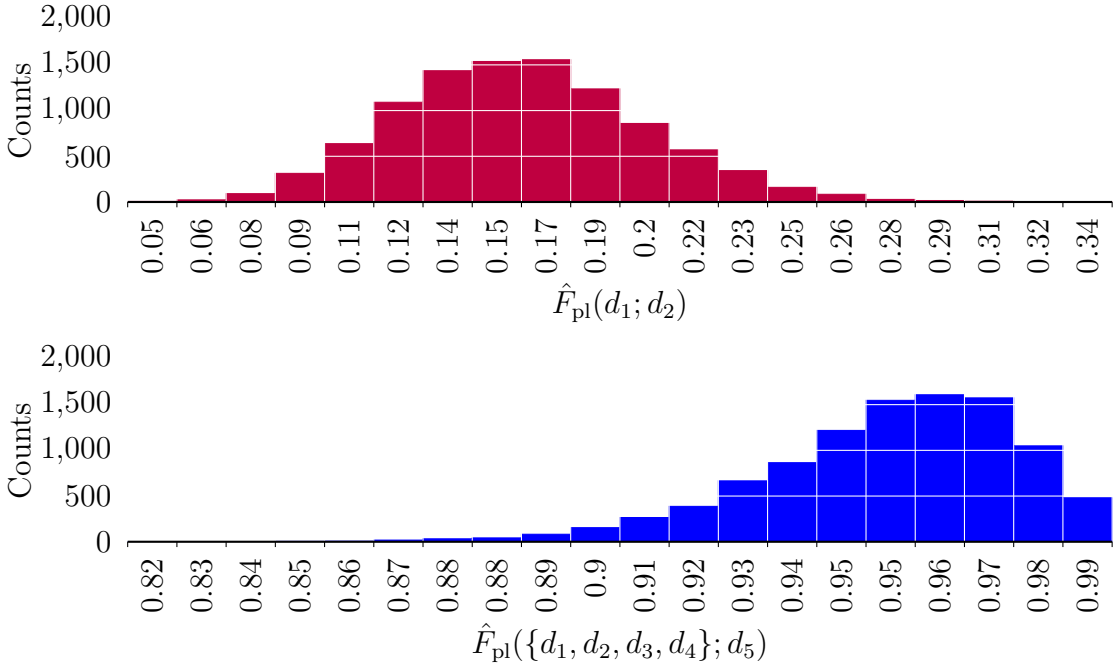


Figure 3.2: Histogram of plugin fraction of information estimates $\hat{F}_{\text{pl}}$ for independent dice rolls. **Top**: for a pair of dice d_1, d_2, we perform 50 independent rolls and compute $\hat{F}_{\text{pl}}(d_1; d_2)$. This process is repeated with 10000 simulations, and we plot the histogram for 20 equal-frequency bins. Despite having a population value of $F(d_1; d_2) = 0$, the histogram has a right-tailed bell shape with expected value 0.17. **Bottom**: same procedure but with 5 dice. Here the histogram has a left tail, with expected value 0.95.

sample of the joint probability distribution $p(\mathcal{I}, Y)$ is a process susceptible to statistical errors. For the discrete data case in particular, the empirical estimator $\hat{I}_{\text{pl}}$, while asymptotically efficient [AK01], exhibits an increasing bias with the domain size of the variables involved that leads to an overestimation of the actual degree of dependency[15] (see Fig. 3.2 for a demonstration). Hence, it is not suited for optimization where we have to soundly compare different variable sets $\mathcal{X} \subseteq \mathcal{I}$ of

[15]Note that this error is not applicable to data management solutions as they operate purely empirically, i.e., they implicitly assume $p = \hat{p}$, while in feature selection it is mitigated due to the low-order approximations (e.g., pairwise). For Markov blanket discovery, however, the independence tests are error prone in a similar manner and remedies include requiring that data samples are at least five times the number of degrees of freedom in the test [PNBT07].

varying dimensionality and consequently of widely varying domain sizes. Even if an appropriate estimator was available, the search space for the optimization problem is exponential in the number of attributes d and exhaustive search is infeasible. At the same time it is unlikely to admit a polynomial time solution [KG05]. To the best of our knowledge, this is the second attempt for a solution to Eq.(3.1). The first attempt by Nguyen et al., [VCB14], while it addresses the need for a correction and proposes a solution based on asymptotics (see Sec. 3.5), the exact search algorithm proposed is only applicable for low-dimensional datasets.[16] In addition, our work makes the connection to causality and focuses on exploratory analysis, while Nguyen et al. consider feature selection. In the next sections we present our solution and the following contributions:

- to correct for overestimation, we derive a consistent and robust estimator for mutual information, as well as accompany it with a set of useful properties that can be used for optimization (Sec. 3.1),
- we show that maximizing the robust estimator is NP-hard, justifying worst-case exponential as well as heuristic optimization algorithms (Sec. 3.2),
- we derive two effective bounding functions for the robust estimator that can be used by algorithms to prune the search space (Sec. 3.3),
- we propose a branch-and-bound algorithm to discover the α-approximate top dependencies for desired approximation guarantee $\alpha \in (0, 1]$, a fast greedy algorithm, as well as a shrink step to remove potentially uniformative variables from the results (Sec. 3.4), and last,
- we perform an extensive evaluation for the estimator, pruning functions, and resulting discovery framework (Sec. 3.5).

We finish with discussion and conclusions in Section 3.6.

[16] The exact search algorithm proposed is essentially exhaustive search with an upper-bound for the maximum search level derived from the corrected estimator proposed. This solution is inefficient as all subsets below the maximum level will be evaluated.

3.1 Permutation mutual information and properties

In this section we derive a corrected estimator for mutual information, as well as properties that can be used for effective optimization.

3.1.1 Permutation mutual information

Intuitively, the reason why $\hat{I}_{\text{pl}}$ is unreliable as an estimator for Eq.(3.1) is that it does not take into account the confidence in the empirical estimates $\hat{H}_{\text{pl}}(Y|\mathcal{X}=\mathbf{x})$ for some $\mathcal{X} \subseteq \mathcal{I}$. For example, in the extreme case where the empirical count $c_{\mathcal{X}}(\mathbf{x})$ is equal to 1, then $c_{\mathcal{X},Y}(\mathbf{x}, y) = 1$ for one value of $y \in V_Y$ and hence $\hat{H}(Y|\mathcal{X}=\mathbf{x})$ is trivially equal to 0 independent of the true distribution p. This case is likely to occur for many of the sampled values for $\mathcal{X}$ if the data size n is small compared to the observed domain of $\mathcal{X}$—even when $F(\mathcal{X};Y) = 0$, which coincides with the highest error, because then $H(Y|\mathcal{X}=\mathbf{x}) = H(Y)$ while $\hat{H}_{\text{pl}}(Y|\mathcal{X}=\mathbf{x}) = 0$.

The tendency for the plugin estimator $\hat{I}_{\text{pl}}$ to overestimate is more formally explained by its bias (see, e.g., [Rou99])

$$\mathbb{E}[\hat{I}_{\text{pl}}(\mathcal{X};Y) - I(\mathcal{X};Y)] \approx \frac{S_{\mathcal{X},Y} - S_{\mathcal{X}} - S_Y + 1}{2n} \ .$$

We see that the bias is independent of the actual distribution p and it depends solely on the domain sizes $S_{\mathcal{X},Y}, S_{\mathcal{X}}, S_Y$ and the number of samples n. The bias is high when the $\mathcal{X}$ and Y samples produce jointly a large domain $V_{\mathcal{X},Y}$ compared to their marginal domains and sample size n, and is at the highest when $\mathcal{X}$ and Y are independent in the underlying distribution p, i.e., when $I(\mathcal{X};Y) = 0$. Regarding the error of the estimator in general, it is known that the bias is the dominating term for high-dimensional $\mathcal{X}$ [JVHW15].

These last observations suggest a correction for the empirical $\hat{I}_{\text{pl}}(\mathcal{X};Y)$ by subtracting its bias assuming independence for $\mathcal{X}$ and Y. A non-parametric choice for this null hypothesis is the **permutation model** [Lan69, p. 214], arriving at

the bias $\mathbb{E}[\hat{I}_{\text{pl}}(\mathcal{X};Y) - I(\mathcal{X};Y) \mid I(\mathcal{X};Y) = 0]$ expressed as the expected value

$$\mathbb{E}_0[\hat{I}_{\text{pl}}(\mathcal{X};Y)] = \frac{1}{n!} \sum_{\sigma \in S_n} \hat{I}_{\text{pl}}(\mathcal{X};Y_\sigma) \ , \tag{3.2}$$

where S_n denotes the symmetric group of $[n]$, i.e., the set of bijections from $[n]$ to $[n]$, and Y_σ denotes the composition of map Y with the permutation $\sigma \in S_n$, i.e., $Y_\sigma(\cdot) = Y(\sigma(\cdot))$. Essentially, Eq.(3.2) is the average empirical mutual information over all possible sample permutations with fixed marginal counts. With this, the **permutation mutual information** is defined as

$$\hat{I}_0(\mathcal{X};Y) = \hat{I}_{\text{pl}}(\mathcal{X};Y) - \mathbb{E}_0[\hat{I}_{\text{pl}}(\mathcal{X};Y)] \ ,$$

and the **permutation fraction of information** as

$$\hat{F}_0(\mathcal{X};Y) = \hat{I}_0(\mathcal{X};Y)/\hat{H}_{\text{pl}}(Y) \ .$$

With this type of correction we achieve the following desired behaviors. First, we arrive at a consistent estimator for I and consequently F. In particular, Nguyen et al. [NEB10] show that $\lim_{n\to\infty} \mathbb{E}_0[\hat{I}_{\text{pl}}(\mathcal{X};Y)] = 0$, and together with the consistency of the plugin $\hat{I}_{\text{pl}}$ [AK01], we have that $\lim_{n\to\infty} \hat{I}_0(\mathcal{X};Y) = I(\mathcal{X};Y)$. Second, we correct the inflated estimates by accounting for the largest possible bias. Third, we incorporate elements from exact significance testing. We know that mutual information indicates independence at population level with the value 0, and hence, by being unbiased under the null hypothesis, we ensure accurate estimates for independence (see Fig. 3.3 for a demonstration). Compared to exact significance tests, this approach can better adapt to the data at hand as it does not require fixed confidence intervals, and while finding the exact probability for the tail of the null distribution is only feasible for small data, we see below that the expected value is much more efficient. Hence, the robust estimator $\hat{I}_0$ is well-suited to control the number of false positive solutions of Eq.(3.1), efficiently, and without the need of any parameters. From here on we abbreviate the **correction terms** $\mathbb{E}_0[\hat{I}_{\text{pl}}(\mathcal{X};Y)]$ as $m_0(\mathcal{X},Y,n)$ and the normalized version

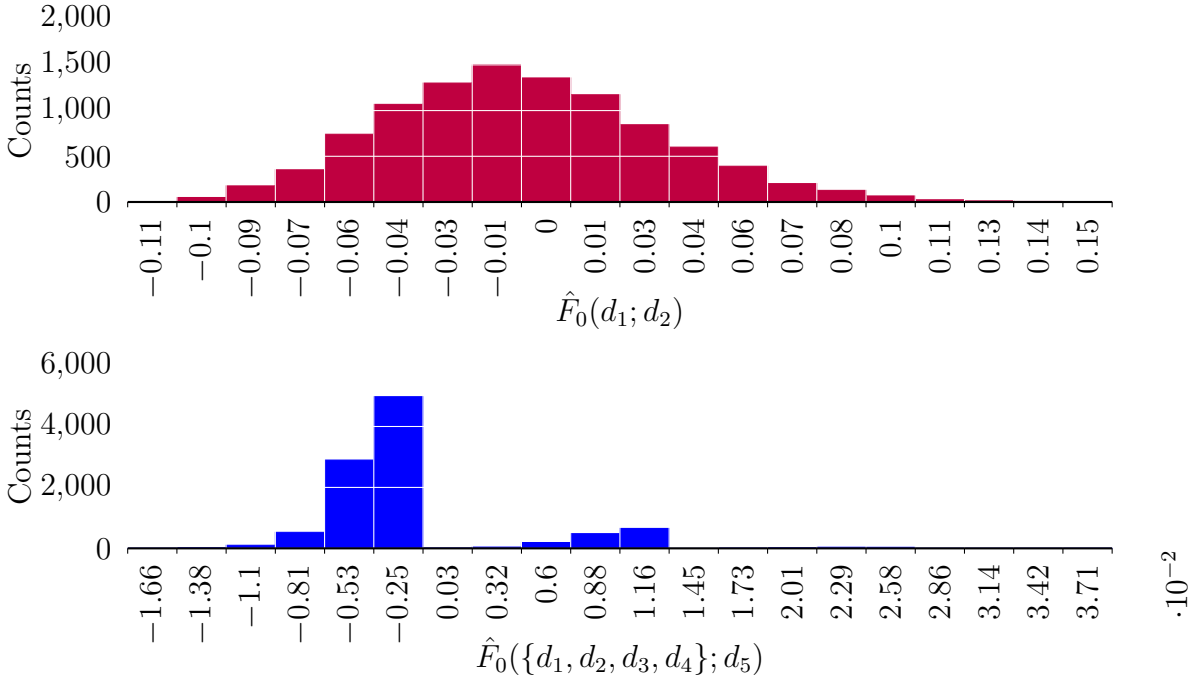


Figure 3.3: Histogram of permutation fraction of information estimates $\hat{F}_0$ for independent dice rolls. **Top**: for a pair of dice d_1, d_2, we perform 50 independent rolls and compute $\hat{F}_0(d_1; d_2)$. This process is repeated with 10000 simulations, and we plot the histogram for 20 equal-frequency bins. **Bottom**: same procedure but with 5 dice. Here, unlike the example in Figure 3.2, the histograms have an expected value of 0. Note that for the 6 dice, the histogram does not have a bell shape because the large domain size of $\{d_1, d_2, d_3, d_4\}$ in combination with the small number of data samples results in most permutations having the same value.

as $b_0(\mathcal{X}, Y, n) = \mathbb{E}_0[\hat{F}_{\text{pl}}(\mathcal{X}; Y)] = m_0(\mathcal{X}, Y, n)/\hat{H}_{\text{pl}}(Y)$.

Regarding the evaluation of Eq.(3.2), a naive approach with $n!$ possible permutations is computationally infeasible. However, Nguyen et al. [NEB09] show that the complexity is dramatically reduced by reformulating it as a function of contingency table cell values and exploiting symmetries. Let the observed domains of $\mathcal{X}$ and Y be $V_{\mathcal{X}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_R\}$ and $V_Y = \{y_1, \ldots, y_C\}$, respectively. We define shortcuts for the observed marginal counts $a_i = c(\mathcal{X} = \mathbf{x}_i)$ and $b_j = c(Y = y_j)$ as well as for the joint counts $c_{i,j} = c(\mathcal{X} = \mathbf{x}_i, Y = y_j)$. The **contingency table c** for $\mathcal{X}$ and Y is then the complete joint count configuration $\mathbf{c} = \{c_{i,j} \colon 1 \leq i \leq R, 1 \leq j \leq C\}$.

The empirical mutual information for $\mathcal{X}$ and Y can then be computed as

$$\hat{I}_{\text{pl}}(\mathcal{X}, Y) = \hat{I}_{\text{pl}}(\mathbf{c}) = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{c_{ij}}{n} \log \frac{c_{ij} n}{a_i b_j} \ .$$

Each $\sigma \in S_n$ results in a contingency table $\mathbf{c}^\sigma$. We denote with $\mathcal{T} = \{\mathbf{c}^\sigma : \sigma \in S_n\}$ the set of all such contingency tables. Crucially, all these tables have the same marginal counts a_i, b_j, $i \in [1, R], j \in [1, C]$. Hence, we can rewrite

$$m_0(\mathcal{X}, Y, n) = \sum_{\mathbf{c} \in \mathcal{T}} \hat{p}_0(\mathbf{c}) \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{c_{ij}}{n} \log \frac{c_{ij} n}{a_i b_j} \ ,$$

where $\hat{p}_0(\mathbf{c})$ is the probability of contingency table $\mathbf{c} \in \mathcal{T}$. This allows us to re-order the terms to have a per-cell contribution to m_0, rather than per-contingency-table $\mathbf{c} \in \mathcal{T}$, i.e.,

$$m_0(\mathcal{X}, Y, n) = \sum_{i=1}^{R} \sum_{j=1}^{C} \sum_{k=0}^{n} \hat{p}_0(c_{ij} = k) \frac{k}{n} \log \frac{kn}{a_i b_j} \ .$$

Under the permutation model, the empirical counts c_{ij} are distributed **hypergeometrically**, i.e.,

$$\hat{p}_0(c_{ij} = k) = \binom{b_i}{k} \binom{n - b_i}{a_j - k} / \binom{n}{a_j} \ .$$

These probabilities can be computed efficiently in an incremental manner using the support of the hypergeometric distribution, i.e., k is non-zero for $k \in [\max(0, a_i + b_j - n), \min(a_i, b_j)]$, and the hypergeometric recurrence formula

$$\hat{p}_0(k+1) = \hat{p}_0(k) \frac{(a_i - k)(b_j - k)}{(k+1)(n - a_i - b_j + k + 1)} \ .$$

The complexity for m_0 is then $O(n \max\{|V_{\mathcal{X}}|, |V_Y|\})$ [RBNV14]. Moreover, the computation can be done in parallel for each individual cell.

In addition to the permutation mutual information and fraction of information,

their **conditional versions** are defined as

$$\hat{I}_0(\mathcal{X}; Y \mid \mathcal{Z}) = \sum_{\mathbf{z} \in V_{\mathcal{Z}}} \hat{p}(\mathbf{z}) \hat{I}_0(\mathcal{X}; Y \mid \mathcal{Z} = \mathbf{z})$$

and

$$\hat{F}_0(\mathcal{X}; Y \mid \mathcal{Z}) = \frac{\hat{I}_0(\mathcal{X}; Y \mid \mathcal{Z})}{\mathbb{E}_0[\hat{H}_{\mathrm{pl}}(Y \mid \mathcal{Z})]} \ ,$$

respectively. Here, $\hat{I}_0(\mathcal{X}; Y \mid \mathcal{Z} = \mathbf{z})$ indicates the permutation mutual information between $\mathcal{X}$ and Y restricted to the data samples $\{i \in [n] : \mathcal{Z}(i) = \mathbf{z}\}$, and $\mathbb{E}_0[\hat{H}_{\mathrm{pl}}(Y \mid \mathcal{Z})] = \hat{H}_{\mathrm{pl}}(Y) - \hat{I}_0(\mathcal{Z}; Y)$ is the conditional Shannon entropy of Y given $\mathcal{Z}$ under the permutation model. Note that we normalize with the corrected conditional entropy, and not with the plugin $\hat{H}_{\mathrm{pl}}(Y \mid \mathcal{Z})$, because otherwise the estimates will be deflated for high-dimensional $\mathcal{Z}$.

In the following section, we couple the above information-theoretic quantities with relations for empirical attributes.

3.1.2 Specializations and Labeling Homomorphisms

Since we identify sets of random variables with their corresponding sample-index-to-value map, they are subject to the following general relations of maps with common domains.

Definition 3.1.1 (**Specialization relation**). *Let A and B be maps defined on a common domain N. We say that A is **equivalent** to B, denoted as $A \equiv B$, if for all $i, j \in N$ it holds that $A(i) = A(j)$ if and only if $B(i) = B(j)$. We say that B is a **specialization** of A, denoted as $A \preceq B$, if for all $i, j \in N$ with $A(i) \neq A(j)$ it holds that $B(i) \neq B(j)$.*

A special case of specializations is given by the subset relation of variable sets, e.g., if $\mathcal{X} \subseteq \mathcal{X}' \subseteq \mathcal{I}$ for some set of variables $\mathcal{I}$, then $\mathcal{X} \preceq \mathcal{X}'$. The specialization relation implies some important properties for empirical probabilities and information-theoretic quantities.

Proposition 3.1.1. *Given variables X, Z, Y, with $X \preceq Z$, the following statements hold:*

a) there is a projection $\pi : V_Z \to V_X$*, s.t. for all* $x \in V_X$*, it holds that* $\hat{p}_X(x) = \sum_{z \in \pi^{-1}(x)} \hat{p}_Z(z)$

b) $\hat{H}_{pl}(X) \leq \hat{H}_{pl}(Z)$

c) $\hat{H}_{pl}(Y \mid Z) \leq \hat{H}_{pl}(Y \mid X)$

d) $\hat{I}_{pl}(X;Y) \leq \hat{I}_{pl}(Z;Y)$

e) $m_0(X,Y,n) \leq m_0(Z,Y,n)$

Proof. Let us denote with p and q the $\hat{p}_{X,Y}$ and $\hat{p}_{Z,Y}$ distributions respectively. Statement (a) follows from the definition. For (b), we define $h(x) = -p(x)\log p(x)$ for $x \in X$, and similarly $h(z)$ for $z \in Z$. We show that for all $x \in X$, $h(x) \leq \sum_{z \in \pi^{-1}(x)} h(z)$. The statement then follows from the definition of $\hat{H}_{\mathrm{pl}}$. We have

$$\begin{aligned} h(x) &= -p(x)\log p(x) \\ &= -\left(\sum_{z \in \pi^{-1}(x)} q(z)\right) \log \left(\sum_{z \in \pi^{-1}(x)} q(z)\right) \\ &= -\sum_{z \in \pi^{-1}(x)} \left(q(z) \log \left(\sum_{s \in \pi^{-1}(x)} q(s)\right)\right) \\ &\leq -\sum_{z \in \pi^{-1}(x)} q(z)\log q(z) = \sum_{z \in \pi^{-1}(x)} h(z) \ , \end{aligned}$$

where the inequality follows from the monotonicity of the log function (and the fact that $q(z)$ is positive for all $z \in Z$).
For (c) let us first recall the log-sum inequality [CT06, p. 31]: for non-negative numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \Big(\sum_{i=1}^{n} a_i\Big) \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \ , \tag{3.3}$$

with equality if and only if a_i/b_i constant. We have

$$\begin{aligned}
\hat{H}_{\mathrm{pl}}(Y \mid Z) &= -\sum_{z\in Z, y\in Y} q(z,y)\log\frac{q(z,y)}{q(z)} \\
&\overset{(a)}{=} -\sum_{x\in X, y\in Y}\sum_{z\in\pi^{-1}(x)} q(z,y)\log\frac{q(z,y)}{q(z)} \\
&\overset{(3.3)}{\leq} -\sum_{x\in X, y\in Y}\Big(\sum_{z\in\pi^{-1}(x)} q(z,y)\Big)\frac{\sum_{z\in\pi^{-1}(x)} q(z,y)}{\sum_{z\in\pi^{-1}(x)} q(z)} \\
&= -\sum_{x\in X, y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)} = \hat{H}_{\mathrm{pl}}(Y \mid X) \ .
\end{aligned}$$

For (d) we have $\hat{I}_{\mathrm{pl}}(Z;Y) = \hat{H}_{\mathrm{pl}}(Y) - \hat{H}_{\mathrm{pl}}(Y \mid Z) \leq \hat{H}_{\mathrm{pl}}(Y) - \hat{H}_{\mathrm{pl}}(Y \mid X) = \hat{I}_{\mathrm{pl}}(X;Y)$ following from (c). For (e), using the chain rule of information and that mutual information is non-negative [CT06, Ch. 2], we have that $\hat{I}_{\mathrm{pl}}(X;Y) \leq \hat{I}_{\mathrm{pl}}(Z;Y)$. Then for each $\sigma \in S_n$ it holds that $\hat{I}_{\mathrm{pl}}(X;Y_\sigma) \leq \hat{I}_{\mathrm{pl}}(Z;Y_\sigma)$, and hence $\sum_{\sigma\in S_n} \hat{I}_{\mathrm{pl}}(X;Y_\sigma) \leq \sum_{\sigma\in S_n} \hat{I}_{\mathrm{pl}}(Z;Y_\sigma)$, which concludes the proof. □

To analyze the monotonicity properties of the permutation model, the following additional definition is useful.

Definition 3.1.2 (**Homomorphic relation**). *We call a labeling X **homomorphic** to a labeling Z w.r.t. to some labeling Y, denoted as $X \precsim_Y Z$, if there exists $\sigma \in S_n$ with $Y \equiv Y_\sigma$ such that $X \preceq Z_\sigma$. We use $X \precsim Z$ whenever clear from the context.*

See Table 3.2 for examples of both introduced relations. Importantly, the inequality of mutual information for specializations carries over to homomorphic variables and in turn to their correction terms.

Proposition 3.1.2. *Given variables X, Z, Y, with $X \precsim Z$, the following statements hold:*

a) $\hat{I}_{pl}(X;Y) \leq \hat{I}_{pl}(Z;Y)$

Table 3.2: Specialization and homomorphism examples. We have $X_1 \preceq X_2$, $X_1 \precsim X_2$, $X_1 \precsim X_3$, $X_1 \precsim X_4$, $X_2 \precsim X_3$. Note that $X_3 \not\precsim X_4$ as there is no $\sigma \in S_4$ that satisfies specialization w.r.t. X_4 and $Y \equiv Y_\sigma$.

X_1	X_2	X_3	X_4	Y
a	a	a	b	a
a	b	b	a	b
b	c	b	b	b
b	c	c	c	b

b) $m_0(X, Y, n) \leq m_0(Z, Y, n)$

Proof. Let $\sigma^* \in S_n$ be a permutation for which $Y \equiv Y_{\sigma^*}$ and $X \preceq Z_{\sigma^*}$. Property (a) follows from

$$\begin{aligned}\hat{I}_{\mathrm{pl}}(Z;Y) &= \hat{I}_{\mathrm{pl}}(Z_{\sigma^*};Y_{\sigma^*}) \\ &= \hat{I}_{\mathrm{pl}}(Z_{\sigma^*};Y) \\ &\geq \hat{I}_{\mathrm{pl}}(X;Y) \ ,\end{aligned}$$

where the inequality holds from Prop. 3.1.1d). For (b), note that for every $\sigma \in S_n$, it holds from Prop. 3.1.1d) that $\hat{I}_{\mathrm{pl}}(Z_{\sigma\circ\sigma^*};Y) \geq \hat{I}_{\mathrm{pl}}(X_\sigma;Y)$. Hence

$$\begin{aligned}m_0(Z,Y,n) =& \frac{1}{n!}\sum_{\sigma\in S_n} \hat{I}_{\mathrm{pl}}(Z_\sigma;Y) \\ =& \frac{1}{n!}\sum_{\sigma\in S_n} \hat{I}_{\mathrm{pl}}(Z_{\sigma\circ\sigma^*};Y) \\ \geq& \frac{1}{n!}\sum_{\sigma\in S_n} \hat{I}_{\mathrm{pl}}(X_\sigma;Y) = m_0(X,Y,n) \ .\end{aligned}$$

□

3.2 Hardness of optimization

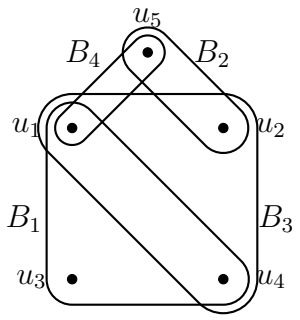
In this section, we prove NP-hardness of maximizing $\hat{F}_0$ (and hence $\hat{I}_0$) by providing a reduction from the well-known NP-hard **minimum set cover** problem: given a finite universe $U = \{u_1, \ldots, u_n\}$ and collection of subsets $\mathcal{B} = \{B_1, \ldots, B_m\} \subseteq 2^U$, find a **set cover**, i.e., a sub-collection $\mathcal{C} \subseteq \mathcal{B}$ with $\bigcup_{B \in \mathcal{C}} B = U$, that is of minimal cardinality [KV12, Ch. 16.1]. A **partial set cover** $\mathcal{C} \subseteq \mathcal{B}$ is one where $\bigcup_{B \in \mathcal{C}} B \neq U$.

The reduction consists of two parts. First, we construct a base transformation $\tau_1(U, \mathcal{B}) = \mathbf{D}_l$ that maps a set cover instance to a dataset $\mathbf{D}_l$, such that a) the plugin $\hat{F}_{\text{pl}}$ is monotonically increasing with coverage, b) set covers correspond to attribute sets with an empirical fraction of information score $\hat{F}_{\text{pl}}$ of 1, and c) correction terms b_0 are a monotonically increasing function of their cardinality. In a second step, we calibrate the b_0 terms such that all candidate set covers have a higher $\hat{F}_0$ value than partial set covers. The latter is achieved by copying the dataset $\mathbf{D}_l$ a suitable number of times k such that the correction terms are sufficiently small but the overall transformation, denoted $\tau_k(U, \mathcal{B}) = \mathbf{D}_{kl}$, is still of polynomial size. Combining these, we arrive at a polynomial time reduction.

The **base transformation** $\tau_1(U, \mathcal{B}) = \mathbf{D}_l$ is defined as follows. The dataset $\mathbf{D}_l$ contains m descriptive attributes $\mathcal{I} = \{X_1, \ldots, X_m\}$ corresponding to the sets of the set cover instance, and a target attribute Y. The sample size is $l = 2n + m + 1$ with a logical partition of the sample into the three regions $S_1 = [1, n]$, $S_2 = [n + 1, 2n]$, and $S_3 = [2n + 1, l]$. The target attribute Y assigns to data points one of three values corresponding to the three parts, i.e., $Y \colon [l] \to \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}$ with

$$Y(j) = \begin{cases} \mathrm{a}, & j \in S_1 \\ \mathrm{b}, & j \in S_2 \\ \mathrm{c}, & j \in S_3 \end{cases},$$

and the descriptive attributes X_i assign up to $n + 3$ distinct values dependending on the set of universe elements covered by set B_i, i.e., $X_i \colon [l] \to \{1, 2, \ldots, n, \mathrm{a}, \mathrm{b}, \mathrm{c}\}$



		$\mathbf{X_1}$	$\mathbf{X_2}$	X_3	X_4	Y
	1	**1**	**a**	1	1	a
	2	**a**	**2**	2	a	a
S_1	3	**3**	**a**	a	a	a
	4	**4**	**a**	4	a	a
	5	**a**	**5**	a	5	a
	6	**a**	**a**	a	a	b
	7	**a**	**a**	a	a	b
S_2	8	**a**	**a**	a	a	b
	9	**a**	**a**	a	a	b
	10	**a**	**a**	a	a	b
	11	**b**	**c**	c	c	c
	12	**c**	**b**	c	c	c
S_3	13	**c**	**c**	b	c	c
	14	**c**	**c**	c	b	c
	15	**c**	**c**	c	c	c

Figure 3.4: Base transformation example. **Left:** a set cover instance $U = \{u_1, \ldots, u_5\}$ and $\mathcal{B} = \{\mathbf{B_1}, \mathbf{B_2}, B_3, B_4\}$. **Right:** the resulting $\mathbf{D}_{15}$ using $\tau_1(U, \mathcal{B})$, with bold indicating the set cover)

with

$$X_i(j) = \begin{cases} j, & j \in S_1 \wedge u_j \in B_i \\ \text{a}, & (j \in S_1 \wedge u_j \notin B_i) \vee j \in S_2 \\ \text{b}, & j = 2n + i \\ \text{c}, & j \in S_3 \setminus \{2n + i\} \end{cases} .$$

See Figure 3.4 for an illustration.

In a nutshell, the base transformation establishes a one-to-one correspondence between $\mathcal{C} \subseteq \mathcal{B}$ and variable sets $\mathcal{X} \subseteq \mathcal{I}$, which we denote with $\mathcal{I}(\mathcal{C})$. We note the following **two remarks**. Let us use **a** for $(\text{a}, \ldots, \text{a})$, and $\bigcup \mathcal{C}$ as a short-cut for $\bigcup_{B \in \mathcal{C}} B$. We have that S_1 and S_2 couple the amount of uncovered elements

$U \setminus \bigcup \mathcal{C}$ to the conditional entropy $\hat{H}_{\mathrm{pl}}(Y \,|\, \mathcal{I}(\mathcal{C}) = \mathbf{a})$ via

$$\hat{p}(Y = \mathrm{a} \,|\, \mathcal{I}(\mathcal{C}) = \mathbf{a}) = |U \setminus \bigcup \mathcal{C}|/(n + |U \setminus \bigcup \mathcal{C}|) \ .$$

In addition, part S_3 links the size of $\mathcal{C}$ to the number of distinct values of $\mathcal{I}(\mathcal{C})$ on S_3, i.e., $|\mathcal{C}| = V_{\mathcal{I}(\mathcal{C})_{S_3}} - 1$. We now establish three central properties for the base transformation.

Lemma 3.2.1. *Let $\tau_1(U, \mathcal{B}) = \mathbf{D}_l$ be the transformation of a set cover instance $(U, \mathcal{B})$, and $\mathcal{C}, \mathcal{C}' \subseteq \mathcal{B}$ two sets. The following statements hold.*

a) If $|\bigcup \mathcal{C}| \geq |\bigcup \mathcal{C}'|$, then $\hat{F}_{pl}(\mathcal{I}(\mathcal{C}); Y) \geq \hat{F}_{pl}(\mathcal{I}(\mathcal{C}'); Y)$, i.e., the plugin $\hat{F}_{pl}$ is monotonically increasing with coverage, and in particular, $\mathcal{C}$ is a set cover if and only if $\hat{F}_{pl}(\mathcal{I}(\mathcal{C}); Y) = 1$,.

b) If C is a set cover and C' is not*, then $\hat{I}_{pl}(\mathcal{I}(\mathcal{C}); Y) - \hat{I}_{pl}(\mathcal{I}(\mathcal{C}'); Y) \geq 2/l$.*

c) If $\mathcal{C}$ and $\mathcal{C}'$ are both set covers, then $\mathcal{I}(\mathcal{C}) \precsim \mathcal{I}(\mathcal{C}')$ if and only if $|\mathcal{C}| \leq |\mathcal{C}'|$.

Proof. Statement (a) follows from the definition of τ_1.

To show (b), since $\hat{F}_{\mathrm{pl}}(\mathcal{I}(\mathcal{C}'); Y)$ and thus $\hat{I}_{\mathrm{pl}}(\mathcal{I}(\mathcal{C}'); Y)$ are monotone in $|\bigcup \mathcal{C}'|$, it is sufficient to consider the case where $|U \setminus \bigcup \mathcal{C}'| = 1$, i.e., only one element $u \in U$ is uncovered. In this case we have

$$\hat{I}_{\mathrm{pl}}(\mathcal{I}(\mathcal{C}); Y) - \hat{I}_{\mathrm{pl}}(\mathcal{I}(\mathcal{C}'); Y) = \hat{H}_{\mathrm{pl}}(Y \,|\, \mathcal{I}(\mathcal{C}')) - \underbrace{\hat{H}_{\mathrm{pl}}(Y \,|\, \mathcal{I}(\mathcal{C}))}_{=0}$$

and, moreover, as required

$$\begin{aligned} \hat{H}_{\mathrm{pl}}(Y \,|\, \mathcal{I}(\mathcal{C}')) &= -\hat{p}(\mathbf{a}, \mathrm{a}) \log \hat{p}(\mathrm{a} \,|\, \mathbf{a}) - \hat{p}(\mathbf{a}, \mathrm{b}) \log \hat{p}(\mathrm{b} \,|\, \mathbf{a}) \\ &= -\frac{1}{l} \log \left(\frac{1}{n+1} \right) - \frac{n}{l} \log \left(\frac{n}{n+1} \right) \geq \frac{2}{l} \ . \end{aligned}$$

For (c) observe that for variable set $\mathcal{X} = \mathcal{I}(\mathcal{C})$ corresponding to set cover $\mathcal{C}$, we have for all $i, j \in S_1$ that $\mathcal{X}(i) \neq \mathcal{X}(j)$. Thus, $\mathcal{X}_{S_1} \equiv \mathcal{X}'_{S_1}$ for variable set

$\mathcal{X}' = \mathcal{I}(\mathcal{C}')$ corresponding to set cover $\mathcal{C}'$. Moreover, we trivially have $\mathcal{X}_{S_2} \equiv \mathcal{X}'_{S_2}$. Finally, let $Q, Q' \subseteq S_3$ denote the indices belonging to S_3 where $\mathcal{X}$ and $\mathcal{X}'$ take on values different from $(c, \ldots, c)$. Note that all values in these sets are unique. Furthermore, if $|\mathcal{C}| \leq |\mathcal{C}'|$ then $|Q| \leq |Q'|$ and in turn $|Q \setminus Q'| \leq |Q' \setminus Q|$. This means we can find a permutation $\sigma \in S_n$ such that for all $i \in Q \setminus Q'$ it holds that $\sigma(i) = j$ with $j \in Q' \setminus Q$ and $\sigma(i) = i$ for $i \notin Q \cap Q'$ (that is σ permutes all indices of non-$(c, \ldots, c)$ values of $\mathcal{C}$ in S_3 to indices of non-$(c, \ldots, c)$ values of $\mathcal{C}'$). For such a permutation it holds that $Y_\sigma \equiv Y$ and $\mathcal{X}_{S_3} \preceq \mathcal{X}'_{S_3\sigma}$. Therefore, $\mathcal{X} \precsim \mathcal{X}'$ as required. □

Now, although set covers $\mathcal{C} \subseteq \mathcal{B}$ correspond to variable sets $\mathcal{I}(\mathcal{X})$ with the maximum empirical fraction of information value of 1, due to the correction term, it can happen that $\hat{F}_0(\mathcal{I}(\mathcal{X}'); Y) \geq \hat{F}_0(\mathcal{I}(\mathcal{X}); Y)$ for a variable set $\mathcal{I}(\mathcal{X}')$ corresponding to a partial set cover. To prevent this, we make use of the following upper-bound of the expected mutual information under the permutation model.

Proposition 3.2.1 ([NEB10], Thm. 7)**.** *For a sample of size n of the joint distribution of variables A and B with domain sizes S_A and S_B respectively, it holds that*

$$m_0(A, B, n) \leq \log\left(\frac{n + S_A S_B - S_A - S_B}{n - 1}\right) .$$

Proposition 3.2.1 implies that we can arbitrarily shrink the correction terms if we increase the sample size but leave the number of distinct values constant. Thus, we define the **extended transformation** $\tau_i(U, \mathcal{B}) = \mathbf{D}_{il}$ through simply copying $\mathbf{D}_l$ a number of i times, i.e., by defining $\mathbf{d}_j = \mathbf{d}_{(j \mod l)}$ for $j \in [l+1, il]$. With this definition, we proceed with the NP-hard result.

Theorem 3.2.1. *Given an i.i.d. sample of the joint distribution of random variables $\mathcal{I} = \{X_1, \ldots, X_d\}$ and Y, the problem of maximizing $\hat{F}_0(\,\cdot\,; Y)$ over all possible subsets $\mathcal{X} \subseteq \mathcal{I}$ is NP-hard.*

Proof. First, let us assume that there exists a number $k \in O(l)$ such that w.r.t. transformation τ_k, all set covers $\mathcal{C} \subseteq \mathcal{B}$ and their corresponding variable sets $\mathcal{X} = \mathcal{I}(\mathcal{C})$ have correction terms with $m_0(\mathcal{X}, Y, kl) < 2/l$. Since all properties of

Lemma 3.2.1 transfer from τ_1 to τ_k, this implies that for all variable sets $\mathcal{X}' = \mathcal{I}(\mathcal{C}')$ corresponding to partial set covers $\mathcal{C}' \subseteq \mathcal{B}$, it holds that

$$\begin{aligned}\hat{F}_0(\mathcal{X};Y) &= \hat{F}_{\mathrm{pl}}(\mathcal{X};Y) - m_0(\mathcal{X},Y,kl)/\hat{H}_{\mathrm{pl}}(Y) \\ &> \hat{F}_{\mathrm{pl}}(\mathcal{X};Y) - 2/(l\hat{H}_{\mathrm{pl}}(Y)) \\ &\geq \hat{F}_{\mathrm{pl}}(\mathcal{X};Y) - (\hat{I}_{\mathrm{pl}}(\mathcal{X};Y) - \hat{I}_{\mathrm{pl}}(\mathcal{X}';Y))/\hat{H}_{\mathrm{pl}}(Y) \\ &= \hat{F}_{\mathrm{pl}}(\mathcal{X}';Y) \geq \hat{F}_0(\mathcal{X}';Y) \ ,\end{aligned}$$

where the greater-than follows from Lemma 3.2.1a) and 3.2.1b). Thus, all $\mathcal{X}$ corresponding to set covers have larger $\hat{F}_0$ than partial set covers. Moreover, we know that $\mathcal{C}$ must be a minimum set cover as required, because for a smaller set cover $\mathcal{C}'$, we would have $\mathcal{I}(\mathcal{C}') \precsim \mathcal{I}(\mathcal{C})$ by Lemma 3.2.1c), and thus $b_0(\mathcal{I}(\mathcal{C}'),Y,kl) \leq b_0(\mathcal{I}(\mathcal{C}),Y,kl)$ from Proposition 3.1.2b). Therefore, $\mathcal{I}(\mathcal{C})$ would not maximize $\hat{F}_0$.

Now, to find the number k that defines the final transformation τ_k, let $\mathbf{D}_{il} = \tau_i(U,\mathcal{B})$ and $\mathcal{C}$ be a set cover of $(U,\mathcal{B})$. Since $\mathcal{X} = \mathcal{I}(\mathcal{C})$ has at most l distinct values in $\mathbf{D}_{il}$ and Y exactly 3, from Proposition 3.2.1 and the monotonicity of ln, we have that

$$\ln(2) m_0(\mathcal{I}(\mathcal{C}),Y,n) \leq \ln\left(\frac{il+3l}{il-1}\right) \leq \ln\left(\frac{i+3}{i-1}\right) \leq \frac{4}{i-1} \ ,$$

where the last inequality follows from $\ln(x) \leq x-1$. Thus, for $k > 2l/\ln 2 + 1 \in O(l)$ we have $m_0(\mathcal{X},Y,kl) < 2/l$ as required. The proof is concluded by noting that the final transformation $\tau_k(U,\mathcal{B})$ is of size $O(l^2 m)$ (where $l = 2n+m+1$), which is polynomial in the size of the set cover instance. □

3.3 Admissible bounding functions for pruning

The NP-hardness established in the previous section excludes the existence of a polynomial time algorithm for maximizing the permutation fraction of information over all $\mathcal{X} \subseteq \mathcal{I}$ (unless P=NP), leaving therefore exact but exponential search and heuristics as the two options. For both, and particularly the former, reducing the search space can lead to more effective algorithms. For this, we derive in this

section bounding functions for the permutation fraction of information $\hat{F}_0$ to be used for pruning.

An **admissible bounding function** $\bar{f}$, also called an **optimistic estimator**, is an upper bound to the optimization function value f over all supersets of a candidate solution $\mathcal{X} \subseteq \mathcal{I}$. The value $\bar{f}(\mathcal{X})$ is called the **potential** of node $\mathcal{X}$, and it must hold that $\bar{f}(\mathcal{X}) \geq f(\mathcal{X}')$ for all $\mathcal{X}'$ with $\mathcal{X} \subseteq \mathcal{X}' \subseteq \mathcal{I}$. With this property, all supersets $\mathcal{X}'$ of $\mathcal{X}$ can be pruned if $\bar{f}(\mathcal{X}) \leq f(\mathcal{X}^*)$, where $\mathcal{X}^*$ is the current best candidate solution found during search. Therefore, for optimal pruning, the bounding function has to be as tight as possible. At the same time, it needs to be efficiently computable. For example, while the **ideal bounding function** for the permutation fraction of information would be

$$\bar{f}_{\text{ideal}}(\mathcal{X}) = \max\{\hat{F}_0(\mathcal{X}'; Y) \colon \mathcal{X} \subseteq \mathcal{X}' \subseteq \mathcal{I}\} \ ,$$

solving it is equivalent to the original problem of Eq.(3.1) and hence NP-hard.

A first attempt for an efficient bounding function involves the upper bound of the fraction of information (i.e., $F = 1$) and the monotonicity of the b_0 term with respect to the subset relation (Prop. 3.1.1e)). In particular, for all $\mathcal{X} \subseteq \mathcal{X}' \subseteq \mathcal{I}$, it follows that

$$\begin{aligned} \hat{F}_0(\mathcal{X}'; Y) =& \hat{F}_{\text{pl}}(\mathcal{X}'; Y) - b_0(\mathcal{X}', Y, n) \\ \leq& 1 - b_0(\mathcal{X}, Y, n) \ . \end{aligned}$$

Hence, we define

$$\bar{f}_{\text{mon}}(\mathcal{X}) = 1 - b_0(\mathcal{X}, Y, n) \tag{3.4}$$

to be the **monotonicity bounding function**. This optimistic estimator is both inexpensive,[17] and applicable to any estimator that has a monotonically increasing correction term. However, it is potentially loose as it assumes that full information about the target can be attained, without the penalty of an increased b_0 term.

[17] One can cache the $b_0(\mathcal{X}, Y, n)$ term required for Eq.(3.4) while computing $\hat{F}_0(\mathcal{X}; Y)$ for a $\mathcal{X} \subseteq \mathcal{I}$ during search.

An alternative idea leading to a more principled admissible bounding function, is to relax the maximum over all supersets to the maximum over all specializations of $\mathcal{X}$. We define the **specialization bounding function** $\bar{f}_{\text{spc}}(\mathcal{X})$ through

$$\begin{aligned}\bar{f}_{\text{spc}}(\mathcal{X}) &= \max\{\hat{F}_0(\mathcal{X}';Y)\colon\ \mathcal{X} \preceq \mathcal{X}'\} \\ &\geq \max\{\hat{F}_0(\mathcal{X}';Y)\colon\ \mathcal{X} \subseteq \mathcal{X}' \subseteq \mathcal{I}\} = \bar{f}_{\text{ideal}}(\mathcal{X})\ .\end{aligned} \tag{3.5}$$

While Eq.(3.5) constitutes an admissible bounding function, it is unclear how it can be efficiently evaluated. To do so, let us denote by R^+ the operation of joining a labeling R with the target attribute Y, i.e., $R^+ = \{R, Y\}$ (see Table 3.3 for an example). This definition gives rise to a simple constructive form for computing $\bar{f}_{\text{spc}}$.

Theorem 3.3.1. *The function $\bar{f}_{spc}$ can be efficiently computed as $\bar{f}_{spc}(\mathcal{X}) = \hat{F}_0(\mathcal{X}^+;Y)$ in time $O(n|V_{\mathcal{X}}||V_Y|)$.*

Proof. We start by showing that the $(\cdot)^+$ operation causes a **positive gain** in $\hat{F}_0$, i.e., for an arbitrary labeling R it holds that $\hat{F}_0(R^+;Y) \geq \hat{F}_0(R;Y)$. It is sufficient to show that $\hat{I}_0(R^+;Y) \geq \hat{I}_0(R;Y)$. We have

$$\begin{aligned}\hat{I}_0(R^+;Y) =& \left(\hat{H}_{\text{pl}}(Y) + \hat{H}_{\text{pl}}(R^+) - \hat{H}_{\text{pl}}(R^+,Y)\right) \\ &-\frac{1}{n!}\left(\sum_{\sigma\in S_n}(\hat{H}_{\text{pl}}(Y_\sigma) + \hat{H}_{\text{pl}}(R^+) - \hat{H}_{\text{pl}}(R^+,Y_\sigma)\right) \\ =&\frac{1}{n!}\sum_{\sigma\in S_n}\hat{H}_{\text{pl}}(R^+,Y_\sigma) - \hat{H}_{\text{pl}}(R^+,Y) \\ \geq&\frac{1}{n!}\sum_{\sigma\in S_n}\hat{H}_{\text{pl}}(R,Y_\sigma) - \hat{H}_{\text{pl}}(R,Y) = \hat{I}_0(R;Y)\ ,\end{aligned}$$

since $\hat{H}_{\text{pl}}(R^+,Y) = \hat{H}_{\text{pl}}(\{R,Y\},Y) = \hat{H}_{\text{pl}}(R,Y)$, and from Proposition 3.1.1b), for every $\sigma \in S_n$, $\hat{H}_{\text{pl}}(R^+,Y_\sigma) \geq \hat{H}_{\text{pl}}(R,Y_\sigma)$.

To conclude, let $\mathcal{Z}$ be an arbitrary specialization of $\mathcal{X}$. We have by definition of

$\mathcal{Z}$ and $\mathcal{Z}^+$, that $\mathcal{X}^+ \preceq \mathcal{Z}^+$. Moreover, $\hat{F}_{\text{pl}}(\,\cdot\,; Y) = \hat{F}_{\text{pl}}(\{\,\cdot\,\} \cup \{Y\}; Y) = 1$. Thus

$$\begin{aligned}\hat{F}_0(\mathcal{X}^+; Y) =& \hat{F}_{\text{pl}}(\mathcal{X}^+; Y) - b_0(\mathcal{X}^+, Y, n)\\ =& 1 - b_0(\mathcal{X}^+, Y, n)\\ \geq& 1 - b_0(\mathcal{Z}^+, Y, n)\\ =& \hat{F}_0(\mathcal{Z}^+; Y) \geq \hat{F}_0(\mathcal{Z}; Y)\ ,\end{aligned}$$

as required. Here, the first inequality follows from Proposition 3.1.1e), the second from the positive gain of $\mathcal{Z}^+$ over $\mathcal{Z}$.

Regarding the complexity for $\bar{f}_{\text{spc}}$, recall that $b_0(\mathcal{X}, Y, n)$ can be computed in time $O(n \max\{S_{\mathcal{X}}, S_Y\})$. The result follows from $S_{\mathcal{X}^+} \leq S_{\mathcal{X}} S_Y$. □

In a nutshell, the operation $(\cdot)^+$ can only increase the $\hat{F}_0$ value, and $\mathcal{X}^+$ constitutes the most efficient specialization of $\mathcal{X}$ in terms of growth in $\hat{F}_{\text{pl}}$ and b_0 (which is not necessarily attainable by a subset of input variables). Note that the $\mathcal{X}^+$ operation is not computed explicitly since it is obtained as the non-zero cell counts of the joint contingency table for $\mathcal{X}$ and Y (which has to be computed for $\hat{F}_0(\mathcal{X}; Y)$ anyway). The following proposition shows that this idea indeed leads to a superior bound compared to $\bar{f}_{\text{mon}}$.

Proposition 3.3.1. *Let $\mathcal{X} \subseteq \mathcal{I}$ and $\Delta = \bar{f}_{mon}(\mathcal{X}) - \bar{f}_{spc}(\mathcal{X})$. The following statements hold:*

a) $\Delta \geq 0$ for all $\mathcal{X} \subseteq \mathcal{I}$, i.e., $\bar{f}_{spc}$ is a tighter bounding function, and

b) there are datasets $\mathbf{D}_{4l}$ for all $l \geq 1$ s.t. $\Delta \in \Omega(1 - \frac{1}{\log 2l})$, i.e., $\bar{f}_{spc}$ has an unbounded pruning potential over $\bar{f}_{mon}$.

Proof. a)

$$\begin{aligned}\bar{f}_{\text{spc}}(\mathcal{X}) =& 1 - b_0(\mathcal{X}^+, Y, n)\\ \leq& 1 - b_0(\mathcal{X}, Y, n) = \bar{f}_{\text{mon}}(\mathcal{X})\ ,\end{aligned}$$

where the inequality holds from Proposition 3.1.1b) and $\mathcal{X} \preceq \mathcal{X}^+$.

b) For $l \geq 1$ we construct a dataset $\mathbf{D}_{4l}$ with two variables $X\colon [4l] \to \{\mathrm{a}, \mathrm{b}\}$ and $Y\colon [4l] \to [2l]$, with

$$X(i) = \begin{cases} \mathrm{a}, & i \mod 2 = 1 \\ \mathrm{b}, & i \mod 2 = 0 \end{cases}$$

and $Y(i) = \lceil i/2 \rceil$ respectively (see Table 3.3). We have

$$\begin{aligned} \Delta &= 1 - b_0(X, Y, 4l) - 1 + \underbrace{b_0(X^+, Y, 4l)}_{=\hat{H}_{\mathrm{pl}}(Y \mid X_\sigma^+)/\hat{H}_{\mathrm{pl}}(Y)=0} \\ &= \frac{1}{n!} \sum_{\sigma \in S_n} \hat{H}_{\mathrm{pl}}(Y \mid X_\sigma)/\hat{H}_{\mathrm{pl}}(Y) \\ &\geq \min_{\sigma \in S_n} \hat{H}_{\mathrm{pl}}(Y \mid X_\sigma)/\hat{H}_{\mathrm{pl}}(Y) \ . \end{aligned}$$

One can show that the minimum of the last step is attained by the permutation $\sigma^* \in S_n$ with

$$\sigma^*(i) = \begin{cases} 2i - 1, & i \in [1, 2l] \\ 4l - 2(4l - i), & i \in [2l + 1, 4l] \end{cases} ,$$

which corresponds to sorting the a and b values of X (see Table 3.3). For this permutation the normalized conditional entropy evaluates to $1 - 1/\log(2l)$ as required. □

Thus, we have established that $\bar{f}_{\mathrm{spc}}$ is tighter than $\bar{f}_{\mathrm{mon}}$, and even that their ratio, and thus the potential for additional pruning, is unbounded.

Regarding their applicability to other mutual information estimators, $\bar{f}_{\mathrm{mon}}$ only requires monotonicity for the correction term, while $\bar{f}_{\mathrm{spc}}$ additionally needs a positive gain w.r.t. to the $(\cdot)^+$ operation. The former is easier to satisfy. Computationally, $\bar{f}_{\mathrm{spc}}(\mathcal{X})$ is more expensive than $\bar{f}_{\mathrm{mon}}(\mathcal{X})$ by a factor of S_Y. In practice one can combine both optimistic estimators in a chain-like manner: first check

Table 3.3: Construction showing the advantage of bound $\bar{f}_{\text{spc}}$ versus $\bar{f}_{\text{mon}}$. We have $\bar{f}_{\text{spc}}(X) = 1 - b_0(X^+, Y, n) = 0$ while $\bar{f}_{\text{mon}}(X) = 1 - b_0(X, Y, n) \geq 1 - 1/\log(n/2)$, i.e., all specializations of X that contain full information about Y are injective (key) maps (see Prop. 3.3.1).

X	Y	X^+	X_{σ^*}
a	1	(a,1)	a
b	1	(b,1)	a
a	2	(a,2)	a
b	2	(b,2)	a
		⋮	

X	Y	X^+	X_{σ^*}
		⋮	
a	2l-1	(a,2l-1)	b
b	2l-1	(b,2l-1)	b
a	2l	(a,2l)	b
b	2l	(b,2l)	b

the pruning condition w.r.t. $\bar{f}_{\text{mon}}$ and only compute $\bar{f}_{\text{spc}}$ if that first check fails. That is, whenever $\bar{f}_{\text{mon}}(\mathcal{X})$ is sufficient to prune a candidate $\mathcal{X}$ we can still do so with the same computational complexity. We refer to this trick as the **chain bounding function** $\bar{f}_{\text{chn}}$. However, the additional evaluation of $\bar{f}_{\text{spc}}(\mathcal{X})$ can be a disadvantage in case it still does not allow to prune. This trade-off is evaluated in Section 3.5.2.

3.4 Optimization algorithms

In this section we provide exact and heuristic optimization algorithms combined with the bounding functions of Section 3.3 to solve Eq.(3.1). In addition, we propose a post-processing shrink step to remove potentially redundant attributes from the solutions based on conditional dependency measures. In the case of a Bayesian network, the heuristic algorithm corresponds to the grow phase for Markov blanket discovery, while the post-processing step corresponds to the shrink phase and is mandatory together with faithfulness to guarantee the discovery of a Markov blanket. Note that we state the top-1 formulation for simplicity, since these algorithms can be trivially extended for top-k by considering a result set of size k. Moreover, we only solve for minimal solutions and not maximal (see Thm. 3.0.1), leaving the latter for future work (see Sec. 3.6.4).

Algorithm 2 OPUS: Given a set of input variables $\mathcal{I}$, function f, bounding function $\bar{f}$, and $\alpha \in (0, 1]$, the algorithm returns the $\mathcal{X}^* \subseteq \mathcal{I}$ satisfying $f(\mathcal{X}^*) \geq \alpha \max\{f(\mathcal{X}') \colon \mathcal{X}' \subseteq \mathcal{I}\}$

```
1: function OPUS(Q, S)
2:     if Q is empty then
3:         return S
4:     else
5:         (X, Z) = pop(Q)
6:         R = {(X ∪ {Z}, Z) : Z ∈ Z}
7:         X* = arg max{f(X') : X' ∈ R ∪ {S}}
8:         R' = {(X', Z) ∈ R : αf̄(X') > f(X*)}
9:         Z' = {Z : (X', Z) ∈ R'}
10:        [(X_1, Z_1), ..., (X_k, Z_k)] = sort(R')
11:        Q' = Q ∪ {(X_i, Z' \ {Z_1, ..., Z_i}) : i ∈ [k]}
12:        return OPUS(Q', X*)
13: X* = OPUS({(∅, I)}, ∅)
```

3.4.1 Exact search

We instantiate the exact search algorithm with the **branch-and-bound (BnB)** algorithm that as the name suggests, consists of two main ingredients: a strategy to explore the search space and a bound for the optimization function at hand to be used for branch pruning (see, e.g., [MS08, Ch. 12.4]). The former is governed by the **refinement operator** (also known as branching operator), a function $r_{\mathcal{I}} \colon 2^{\mathcal{I}} \to 2^{\mathcal{I}}$ that non-redundantly generates the search space of $\mathcal{I} = \{X_1, \ldots, X_d\}$ from the designated root element $\emptyset$, i.e., for all $\mathcal{X} \in 2^{\mathcal{I}}$ there must be a unique sequence $\emptyset = \mathcal{X}_1, \ldots, \mathcal{X}_l = \mathcal{X}$ such that $\mathcal{X}_{i+1} \in r_{\mathcal{I}}(\mathcal{X}_i)$ for $i = 1, \ldots, l-1$. For example, one such operator is the **alphabetical refinement operator**

$$r^A_{\mathcal{I}}(\mathcal{X}) = \{\mathcal{X} \cup \{X_i\} \colon i > \max\{j \colon X_j \in \mathcal{X}\}, X_i \in \mathcal{I}\} \ . \tag{3.6}$$

Besides being very effective in practice for hard problems, this style of optimization also provides the option of relaxing the required result guarantee to that of an α-approximation for accuracy parameter $\alpha \in (0, 1]$. Hence, using α-values of less than 1 allows to trade accuracy for computation time in a principled

manner. Here, we consider **optimized pruning for unordered search (OPUS)** by Webb [Web95], an advanced variant of branch-and-bound that effectively propagates pruning information to siblings in the search tree. Algorithm 2 shows the details of this approach.

In addition to keeping track of the best solution $\mathcal{X}^*$ explored so far, the algorithm maintains a priority queue $\mathbf{Q}$ of pairs $(\mathcal{X}, \mathcal{Z})$, where $\mathcal{X} \subseteq \mathcal{I}$ is a candidate solution and $\mathcal{Z} \subseteq \mathcal{I}$ constitutes the variables that can still be used to refine $\mathcal{X}$, e.g., $\mathcal{X}' = \mathcal{X} \cup \{Z\}$ for a $Z \in \mathcal{Z}$. The top element is the one with the smallest cardinality and the highest potential $\bar{f}$ (a combination of breadth-first and best-first order). Starting with $\mathbf{Q} = \{(\emptyset, \mathcal{I})\}$, $\mathcal{X}^* = \emptyset$, and a desired approximation guarantee $\alpha \in (0, 1]$, in every iteration Opus creates all refinements of the top element of $\mathbf{Q}$ and updates $\mathcal{X}^*$ accordingly (lines 5-7). Next the refinements are pruned using $\bar{f}$ and α (line 8). Following, the pruned list is sorted according to decreasing potential,[18] the possible refinement elements $\mathcal{Z}'$ are non-redundantly propagated to the refinements of the top element, and finally the priority queue is updated with the new candidates (lines 9-11).

3.4.2 Heuristic search

A commonly used alternative to exponential search for optimizing dependency measures is the standard **greedy algorithm (Grd)**. This algorithm only refines the best candidate in a given iteration. Moreover, bounding functions can be incorporated as an early termination criterion. For the permutation fraction of information in particular, there is potential to prune many of the higher levels of the search space. The algorithm is presented in Algorithm 3.

The algorithm keeps track of the best solution $\mathcal{X}^*$ explored, as well as the best candidate for refinement $\mathcal{C}^*$. Starting with $\mathcal{X}^* = \emptyset$ and $\mathcal{C}^* = \emptyset$, the algorithm in each iteration (i.e., search space level) checks whether $\mathcal{C}^*$ can be refined further, i.e., if $\mathcal{I} \setminus \mathcal{C}^*$ is not empty, or if $\mathcal{C}^*$ has potential (the early termination criterion). If not, the algorithm terminates returning $\mathcal{X}^*$ (lines 2-3). Otherwise $\mathcal{C}^*$ is refined

[18] An admissible heuristic that propagates the most refinement elements to the least promising candidates which have higher chances of being pruned [Web95].

Algorithm 3 GRD: Given a set of input variables $\mathcal{I}$, function f, and bounding function $\bar{f}$, the algorithm returns the $\mathcal{X}^* \subseteq \mathcal{I}$ approximating $f(\mathcal{X}^*) = \max\{f(\mathcal{X}') \colon \mathcal{X}' \subseteq \mathcal{I}\}$

```
1: function GRD(C, S)
2:     if I \ C is empty or f̄(C) ≤ f(S) then
3:         return S
4:     else
5:         R = {C ∪ {Z} : Z ∈ I \ C}
6:         C* = arg max{f(X') : X' ∈ R}
7:         X* = arg max{f(X') : X' ∈ {S, C*}}
8:         return GRD(C*, X*)
9: X* = GRD(∅, ∅)
```

to all possible refinements, and the best one is selected as a candidate to update $\mathcal{X}^*$ (lines 5-7).

Concerning the approximation ratio of the greedy algorithm, there exists a large amount of research focused on submodular and/or monotone functions (see, e.g., [FMV07]). Recall that for a set $\mathcal{I} = \{X_1, \ldots, X_d\}$, a function $f \colon 2^{\mathcal{I}} \to \mathbb{R}$ is called **submodular** if for every $\mathcal{X} \subseteq \mathcal{X}' \subseteq \mathcal{I}$ and $X_i \in \mathcal{I} \setminus \mathcal{X}'$, it holds that

$$f(\mathcal{X}' \cup \{X_i\}) - f(\mathcal{X}') \leq f(\mathcal{X} \cup \{X_i\}) - f(\mathcal{X}) \ ,$$

i.e., it satisfies the **diminishing returns property**. The following proposition establishes that I, $\hat{I}_{\text{pl}}$, and $\hat{I}_0$, are all violating this property.

Proposition 3.4.1. *Given $\mathcal{I} = \{X_1, \ldots, X_d\}$ and target variable Y, the mutual information $I(.;Y)$, the plug-in $\hat{I}_{pl}(.;Y)$, and the permutation $\hat{I}_0(.;Y)$ are not submodular w.r.t. the first argument.*

Proof. We prove it via an intuitive counter example. Let us consider the data of Table 3.4 and the corresponding induced empirical distribution $\hat{p}$. Here B and C are connected to Y via a XOR function, where Y is marginally independent of B and C, but functionally dependent on $\{B, C\}$. For sets $\{A\}$, $\{A, B\}$, and element

Table 3.4: Example data used in Proposition 3.4.1 to show non-submodularity of $I, \hat{I}_{\text{pl}}, \hat{I}_0$.

A	B	C	Y
a	a	a	a
a	a	b	b
a	b	b	a
b	b	a	b

C, we have that

$$\begin{aligned}\hat{I}_{\text{pl}}(\{A,B,C\};Y) - \hat{I}_{\text{pl}}(\{A,B\};Y) =& 0.5 \\ >& \hat{I}_{\text{pl}}(\{A,C\};Y) - \hat{I}_{\text{pl}}(\{A\};Y) = 0.19 \ ,\end{aligned}$$

i.e., there is a violation of the diminishing returns property, and hence $\hat{I}_{\text{pl}}$ is not submodular. By considering $p = \hat{p}$, it is straightforward to show that I is also not submodular.

Regarding $\hat{I}_0$, we have that

$$\begin{aligned}\hat{I}_0(\{A,B,C\};Y) - \hat{I}_0(\{A,B\};Y) =& 0.17 \\ >& \hat{I}_0(\{A,C\};Y) - \hat{I}_0(\{A\};Y) = -0.17 \ ,\end{aligned}$$

and hence $\hat{I}_0$ is not submodular. Also note that while both $\hat{I}_{\text{pl}}$ and I are monotone functions with respect to the subset relation, $\hat{I}_0$ is not because both $\hat{I}_{\text{pl}}$ and the correction b_0 are monotonically increasing (Prop. 3.1.1). □

While approximation results for submodular and/or monotone functions are not directly applicable to $\hat{I}_0$, we empirically evaluate the quality of solutions in Section 3.5.2.

3.4.3 Shrinking

After obtaining the solution $\mathcal{X}^*$ using either the exact or greedy algorithm, one can quantify the marginal gains $Q(X; Y \,|\, \mathcal{X}^* \setminus \{X\})$ for each $X \in \mathcal{X}^*$ for some

Algorithm 4 SHRK: Given solution set $\mathcal{X}^*$, target Y, conditional dependency score Q, and threshold $\phi \in [0, 1]$, the algorithm shrinks the result $\mathcal{X}^*$ according to Q and ϕ

```
1: function SHRK(X*, Y, Q, τ)
2:     while X* does not change do
3:         X = arg min{Q(X; Y | X* \ {X}) : X ∈ X*}
4:         if Q(X; Y | X* \ {X}) ≤ φ then
5:             X* = X* \ {X}
6:     return X*
```

conditional dependency measure Q, and assess the individual contributions to the solution. In Algorithm 4 we present the **shrink step (SHRK)** that removes attributes from the solution if the marginal gain is less then some threshold $\phi \in [0, 1]$. Note that while this procedure can be used optionally to remove low marginally scoring attributes, assuming a Bayesian network and employing the greedy makes it mandatory to guarantee the discovery of a Markov blanket. In fact, the greedy algorithm and shrink step combination is identical to the grow-shrink type of Markov blanket discovery algorithms such as IAMB (Alg. 1), since the conditional fraction of information can be used as a conditional independence test with ϕ regulating the level of conservatism.

3.5 Evaluation

In this section, we empirically evaluate the performance of discovering functional dependencies with the permutation fraction of information $\hat{F}_0$, including the bias and variance of $\hat{F}_0$ as an estimator, the performance of the bounding functions for both branch-and-bound and greedy search, precision and recall on Market blanket discovery, as well as perform qualitative experiments with two case studies.

3.5.1 Estimator performance

Here, we evaluate the **estimated bias** and **variance** of $\hat{F}_0$ for various degrees of dependency. We do so by creating synthetic data from various models for which we know the true fraction of information F. Let us denote by $\mathcal{P}$ the set

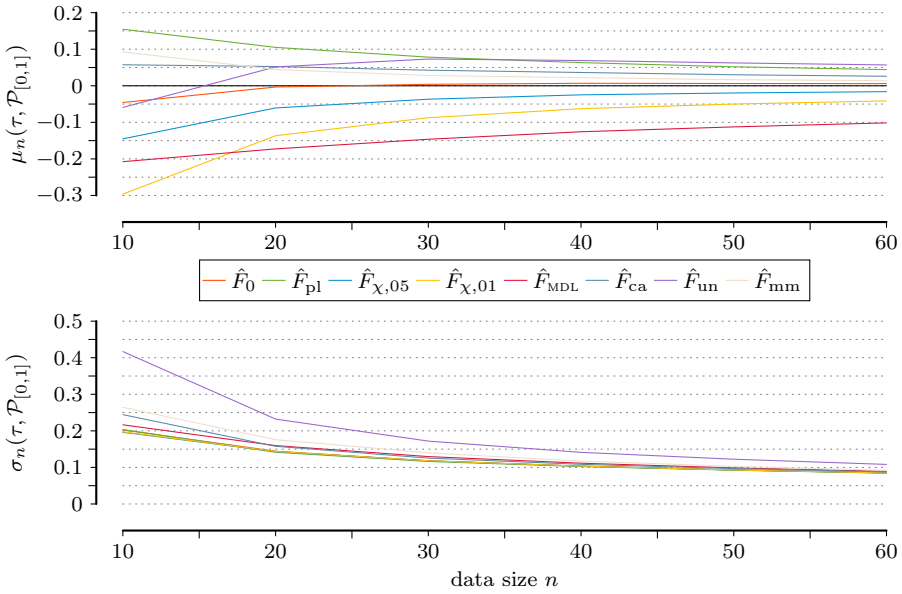


Figure 3.5: Average empirical bias and standard deviation of estimators over all dependency degrees and varying number of samples. Average bias $\mu_n(\tau, \mathcal{P}_{[0,1]})$ (**top**) and average standard deviation $\sigma_n(\tau, \mathcal{P}_{[0,1]})$ (**bottom**) of estimators τ for all 100 sampled pmfs $p^{(i)} \in \mathcal{P}_{[0,1]}$ across data sizes $n = \{10, 20, \ldots, 60\}$.

of all joint probability mass functions over two random variables X and Y with $S_X = S_Y = 3$, and by $\mathcal{P}_{[a,b]}$ all such probability mass functions for which we have a score of $F(X;Y) \in [a, b]$. We consider four dependency regions: weak $\mathcal{P}_{[0,0.25)}$, low $\mathcal{P}_{[0.25,0.5)}$, high $\mathcal{P}_{[0.5,0.75)}$, and strong $\mathcal{P}_{[0.75,1]}$.

We sample uniformly 100 pmfs $p^{(1)}, \ldots, p^{(100)}$, 25 from each dependency region. Note that for every $p^{(i)}$ we know the true $F_{p^{(i)}}$ value and we can sample data $D_n \sim p^{(i)}$ of arbitrary size n. Let $\tau(\mathbf{D}_n)$ be the result of an estimator τ computed on data $\mathbf{D}_n$. We denote with $b_n(p, \tau)$ and $std_n(p, \tau)$ the bias and standard deviation of τ when fixing the underlying pmf to a $p \in \mathcal{P}$, i.e., $b_n(p, \tau) = \mathbb{E}_{\mathbf{D}_n \sim p}[\tau(\mathbf{D}_n)] - F_p$ and $std_n(p, \tau) = \sqrt{\mathbb{E}_{\mathbf{D}_n \sim p}[(\tau(\mathbf{D}_n) - \mathbb{E}_{\mathbf{D}_n \sim p}[\tau(\mathbf{D}_n)])^2]}$. The expectation terms are estimated by sampling per pmf $p^{(i)}$ and n a total of 500 datasets. We average over $\mathcal{P}_{[a,b]}$ regions and end up with estimates $\mu_n(\tau, \mathcal{P}_{[a,b]})$ and $\sigma_n(\tau, \mathcal{P}_{[a,b]})$ for the average bias and standard deviation of estimator τ and sample size n.

For this experiment we evaluate over the different samples sizes $n \in \{10, 20, \ldots, 60\}$

and the following estimators: the plugin $\hat{F}_{\text{pl}}$, the permutation $\hat{F}_0$, the unseen $\hat{F}_{\text{un}}$, the minimax $\hat{F}_{\text{mm}}$ (Sec. 2.2), and in addition, we consider 3 estimators that specifically aim to correct for spurious dependencies. The first, proposed by Nguyen et al. [VCB14], is based on the same correction principle using asymptotics, and particular the chi-square distribution. This corrected estimator, which we term the **chi-square estimator** and denote as $\hat{F}_{\chi,\alpha}$, is defined as

$$\hat{F}_{\chi,\alpha}(\mathcal{X};Y) = \frac{\hat{I}_{\text{pl}}(\mathcal{X},Y) - \frac{1}{2n}\chi_{\alpha,l(\mathcal{X},Y)}}{\hat{H}_{\text{pl}}(Y)} \ ,$$

where $\chi_{\alpha,l(\mathcal{X},Y)}$ is the critical value corresponding to a confidence level $1-\alpha$ and degrees of freedom $l(\mathcal{X},Y) = (\prod_{X\in\mathcal{X}} V_X - 1)(V_Y - 1)$. Here, α controls the amount of penalty with suggested values 0.01 and 0.05. The second by Suzuki [Suz16], which we term the **MDL estimator**, penalizes by the minimum description length principle and is defined as

$$\hat{F}_{\text{MDL}}(\mathcal{X};Y) = \frac{\hat{I}_{\text{pl}}(\mathcal{X};Y) - \frac{1}{2n}l(\mathcal{X},Y)\log(n)}{\hat{H}_{\text{pl}}(Y)} \ .$$

The third follows a similar correction resulting from the application of the quantification adjustment framework proposed by Romano et al. [RVBV16]. We term this estimator the **chance-adjusted estimator** and is defined as

$$\hat{F}_{\text{ca}}(\mathcal{X};Y) = \frac{\hat{I}_{\text{pl}}(\mathcal{X},Y) - \mathbb{E}_0[\hat{I}_{\text{pl}}(\mathcal{X},Y)]}{\hat{H}_{\text{pl}}(Y) - \mathbb{E}_0[\hat{H}_{\text{pl}}(\mathcal{X},Y)]} \ .$$

Note that estimators $\hat{F}_0, \hat{F}_{\text{pl}}, \hat{F}_{\chi,\alpha}, \hat{F}_{\text{MDL}}$ are normalized by the plugin entropy estimator, while $\hat{F}_{\text{un}}, \hat{F}_{\text{mm}}, \hat{F}_{\text{ca}}$ by the entropy estimator corresponding to the style of correction used.

We first focus on the general behavior of the bias and standard deviation for each estimator τ, and plot in Figure 3.5 the average bias $\mu_n(\tau, \mathcal{P}_{[0,1]})$ (top) and average standard deviation $\sigma_n(\tau, \mathcal{P}_{[0,1]})$ (bottom). We observe that the plugin $\hat{F}_{\text{pl}}$ and minimax $\hat{F}_{\text{mm}}$ experience positive bias, with $\hat{F}_{\text{pl}}$ having the largest, as expected. The unseen $\hat{F}_{\text{un}}$ starts with a negative bias for sample size 10, which

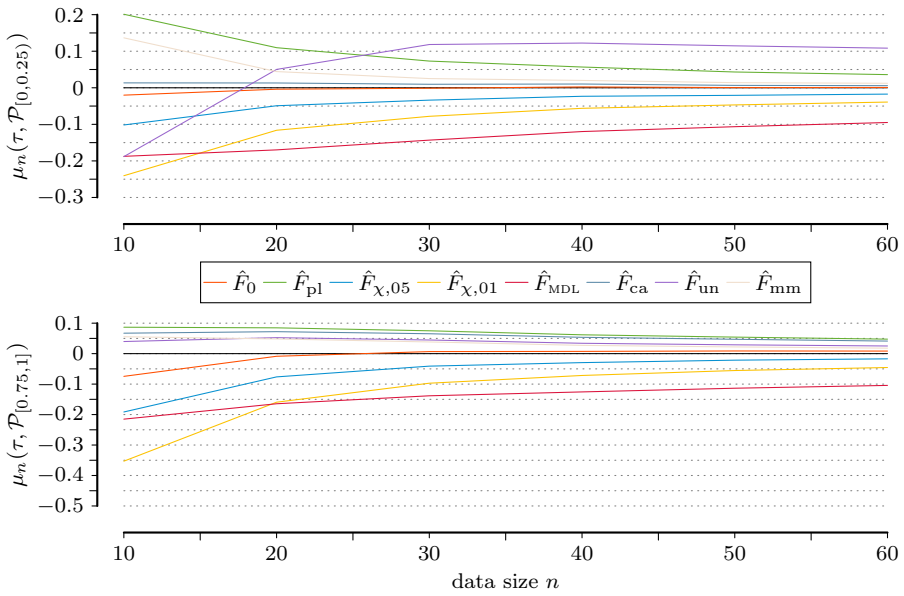


Figure 3.6: Bias of estimators averaged over weak and strong dependencies for varying number of samples. Average bias over all weak dependency $p^{(i)}$ (**top**), and over all strong dependency $p^{(i)}$ (**bottom**) of estimators τ for all 100 sampled pmfs $p^{(i)} \in \mathcal{P}_{[0,1]}$ across data sizes $n = \{10, 20, \ldots, 60\}$.

then turns positive. The chance-adjusted $\hat{F}_{\text{ca}}$ experiences a small positive bias. The remaining estimators all have a negative bias, with $\hat{F}_{\chi,01}$ having the largest. The $\hat{F}_{\text{MDL}}$ also has a large negative bias, with $\hat{F}_{\chi,05}$ following. Notice how $\hat{F}_{\chi,\alpha}$ increases the bias for smaller α. The permutation $\hat{F}_0$ has the smallest negative bias. Regarding the convergence to 0 bias, additive smoothing and MDL are the slowest, with the remaining having good speed, and particularly the permutation $\hat{F}_0$. As for the standard deviation, the unseen $\hat{F}_{\text{un}}$ has the largest by far. The remaining all show similar behavior, with the minimax $\hat{F}_{\text{mm}}$ and chance-adjusted $\hat{F}_{\text{ca}}$ having slightly higher.

It is also informative to consider the bias behavior not on average for all dependency degrees, but specifically for weak and strong dependencies, i.e., the cases closer to independence and functional dependency, respectively. For this, we plot in Figure 3.6 the average biases $\mu_n(\tau, \mathcal{P}_{[0,0.25)})$ (top) and $\mu_n(\tau, \mathcal{P}_{[0.75,1]})$ (bottom). In Figure 3.5, the positive bias estimators (i.e., $\hat{F}_{\text{mm}}$ and $\hat{F}_{\text{pl}}$) have a

large bias for weak dependencies and small bias for strong dependencies. The situation is reverted for the negative bias estimators. It is interesting to note that the unseen $\hat{F}_{\text{un}}$ has for weak dependencies both a large negative bias for $n = 10$, and then a large positive bias for $n > 10$. However, for strong dependencies it has a small positive bias.

We observe in general that the permutation estimator $\hat{F}_0$ has a consistent behavior with very small negative bias across all degrees of dependency, with comparable standard deviation and fast convergence. Note that this behavior is obtained without any parameter, unlike $\hat{F}_{\chi,\alpha}$, but rather it adapts to the data at hand by employing the data-dependent expected value under the permutation model. Hence, it is well-suited for exploratory tasks.

3.5.2 Optimization performance

We next investigate the optimization performance of the algorithms and bounding functions proposed on real-word data. Our code is available online.[19]

We consider datasets from the KEEL data repository [SRAFFH+11]. In particular, we use all classification datasets with $d \in [10, 90]$ and no missing values, resulting in 35 datasets with 52000 and 30 rows and columns on average, respectively. All metric attributes are discretized in 5 equal-frequency bins. The datasets are summarized in Table 3.5. The runtimes are averaged over 3 runs.

We use two metrics for evaluation, the **relative runtime difference (rrd)** and the **relative difference in number of explored nodes (rnd)**. For methods A and B, the relative runtime difference on a particular dataset is computed as

$$\text{rrd}(A, B) = \frac{(\tau_A - \tau_B)}{\max\{\tau_A, \tau_B\}} \ ,$$

where τ_A and τ_B are the run times for A and B respectively. The rrd score lies in $[-1, 1]$, where positive (negative) values indicate that B is proportionally faster (slower). For example, a rrd score of 0.5 corresponds to a factor of 2 speed-up, 0.66 to a factor of 3, 0.75 to 4 etc. The relative nodes explored difference rnd

[19] https://github.com/pmandros/fodiscovery

is defined similarly. For both scores, we consider $(-0.5, 0.5)$ to be a region of practical equivalence, i.e., a factor of 2 of improvement is required to consider a method "better".

Branch-and-bound

We first investigate the performance of the branch-and-bound algorithm. We consider OPUS$_{\text{chn}}$ and OPUS$_{\text{mon}}$, i.e., Algorithm 2 with the chain bounding function $\bar{f}_{\text{chn}}$ (i.e., $\bar{f}_{\text{spc}}$ and $\bar{f}_{\text{mon}}$ combined) and the monotonicity $\bar{f}_{\text{mon}}$, respectively. For a fair comparison, we set a common α value for both methods on each dataset by determining the largest α value in increments of 0.05 such that they terminate in less than 90 minutes. The results are in Table 3.5. Regarding runtime, OPUS$_{\text{chn}}$ and OPUS$_{\text{mon}}$ require 296 and 360 seconds on average, respectively. For the majority of the data, both need less than 10 minutes. The approximation guarantees α are 0.85 on average, with 23 out of 35 datasets having $\alpha = 1$, i.e., an optimal solution.

For a more thorough comparison, in Figure 3.7 we present the rnd and rrd for OPUS$_{\text{chn}}$ and OPUS$_{\text{mon}}$. The top plot demonstrates that $\bar{f}_{\text{spc}}$ can lead to a considerable reduction of nodes explored over $\bar{f}_{\text{mon}}$, that in absolute numbers comes down to roughly 50% on average (41434 versus 78309). More specifically, 15 cases have at least a factor of 2 reduction, 7 have 4, and there is one 1 with 760. For 20 cases there is no practical difference. The plot validates that the potential for additional pruning is indeed unbounded (Sec. 3.3). In terms of runtime efficiency (bottom), OPUS$_{\text{chn}}$ is "faster" in 70% of the datasets. In more detail, and considering practical improvements, 12 datasets have at least a factor of 2 speedup, 6 have 4, 1 has 266, while only 2 have a factor of 2 slowdown. Moreover, we observe from the plot (since datasets are sorted in decreasing number of attributes) a clear correlation between number of attributes and efficiency: the 6 out of 10 datasets with the slowdown are also the ones with the lowest number of features. We observe in general that both bounding functions, and particularly the $\bar{f}_{\text{spc}}$, make the branch-and-bound search very effective in practice.

In Table 3.5 we also report the maximum depth and solution depth for OPUS$_{\text{chn}}$, i.e., the maximum level of the search space the algorithm had to explore and in

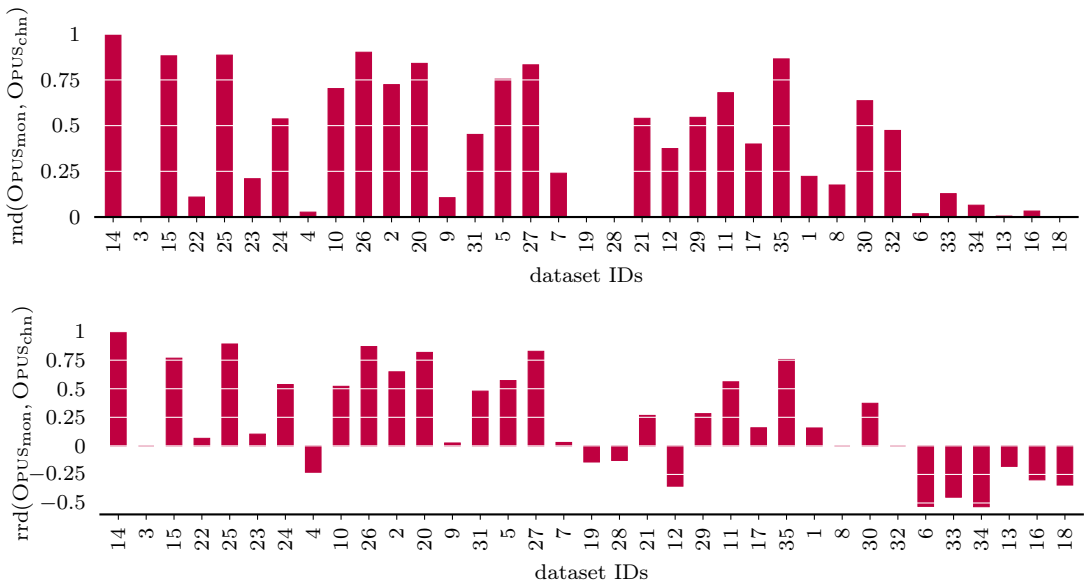


Figure 3.7: Evaluating the bounding functions for branch-and-bound optimization. Relative nodes explored difference (top) and relative runtime difference (bottom) between methods OPUS_{chn} and OPUS_{mon}. Positive (negative) numbers indicate that OPUS_{chn} (OPUS_{mon}) is proportionally more effective. The datasets are sorted in decreasing number of attributes.

which level the solution was found. First, we see that $\hat{F}_0$ retrieves solutions small in cardinality, 3.6 on average, which is a reasonable number for the size of the data considered. We also see that $\bar{f}_{\text{chn}}$ with 5.9 maximum depth level on average, prunes many of the higher levels of the search space, which explains to a large extend the effectiveness of OPUS.

GREEDY

We now proceed with the evaluation for the heuristic search. We present the relative runtime differences of GRD and GRD_{chn}, i.e., Algorithm 3 with and without $\bar{f}_{\text{chn}}$, in Figure 3.8 (results in Tab. 3.5). While the greedy algorithm is fast with 32 and 51 seconds on average with and without pruning, respectively, the plot shows that $\bar{f}_{\text{chn}}$ indeed improves the efficiency of the heuristic search, as we find that for 12 datasets there is a speedup of at least a factor of 2, and 8 of at least a factor of 4.

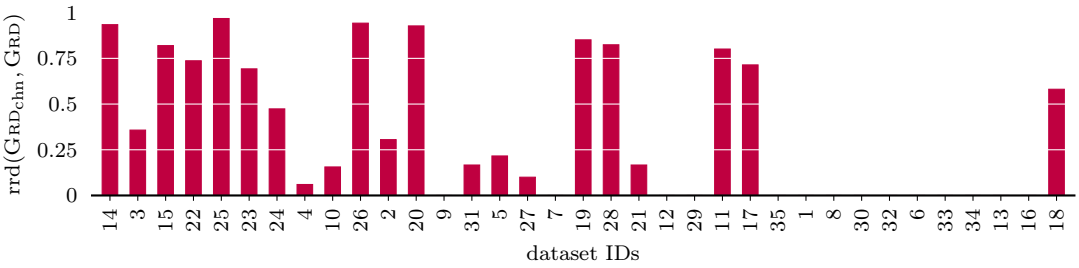


Figure 3.8: Evaluating $\bar{f}_{\mathbf{spc}}$ for heuristic optimization. Relative time difference between methods GRD_{chn} and GRD. Positive (negative) numbers indicate that GRD_{chn} (GRD) is proportionally more effective. The datasets are sorted in decreasing number of attributes.

Next, we investigate the quality of the greedy results. Note that this is possible as we have access to the branch-and-bound results. In Figure 3.9 we plot the differences between the $\hat{F}_0$ score of the results obtained by greedy and branch-and-bound on each dataset. Note that branch-and-bound uses the same α values as with the experiments in Section 3.5.2, and that we only show the non-zero solution differences: left for $\alpha = 1$, i.e, optimal solutions, and right for $\alpha < 1$, i.e., approximate solutions with guarantees. We observe that there is no difference in 21 out of 35 cases considered, 7 where greedy is better,[20] and 7 where branch-and-bound is better. Out of the 21 cases where the two algorithms have equal $\hat{F}_0$, 16 of them have $\alpha = 1$, i.e., the greedy algorithm is optimal roughly 45% of the time. Moreover, the cases where branch-and-bound is better is only by a small margin, 0.03 on average, while greedy "wins" by 0.1 on average. Another observation from the right plot of Figure 3.9 is that the largest differences are for the 3 datasets where the lowest α values where used, i.e., $0.05, 0.1$, and 0.35.

In Figure 3.10 we consider the relative runtime difference between greedy and branch-and-bound, i.e., GRD_{chn} and OPUS_{chn}. As expected, the greedy algorithm is significantly faster in the majority of cases. There are, however, 4 cases where branch-and-bound terminates much faster, which also happen to coincide with more aggressive α values.

These results suggest that heuristic optimization with Algorithm 3 is a good

[20]This of course on the datasets for which $\alpha < 1$.

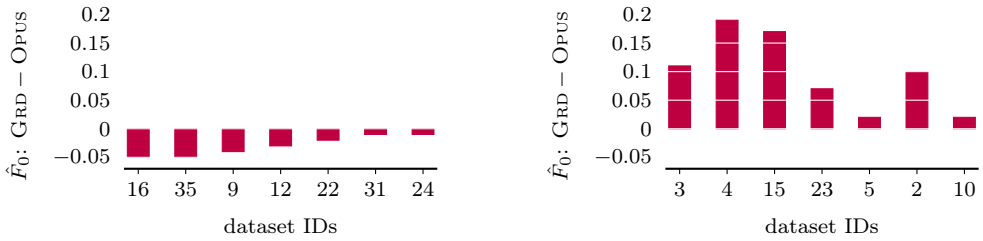


Figure 3.9: Evaluating the heuristic algorithm for result quality. Left: difference in $\hat{F}_0$ between Grd and Opus solutions (i.e., $\hat{F}_0(\mathcal{X}^*_{grd}; Y) - \hat{F}_0(\mathcal{X}^*_{bnb}; Y)$ where $\mathcal{X}^*_{grd}$ and $\mathcal{X}^*_{bnb}$ are the solutions of Algorithm 3 and 2 respectively) for $\alpha = 1$. Since $\alpha = 1$, the negative values close to 0 indicate that Alg. 3 retrieves nearly optimal solutions. Data are sorted in increasing quality difference. **Right**: difference for $\alpha < 1$. Positive values indicate that Alg. 3 retrieves better solutions when Alg. 2 uses guarantees $\alpha < 1$. Data are sorted in increasing α values.

option for the permutation fraction of information as it produces fast and nearly optimal solutions.

3.5.3 Markov blanket discovery on Bayesian networks

Next we evaluate the algorithms and estimators proposed for maximizing the fraction of information on the task of Markov blanket discovery. For this, we use the Alarm dataset [BSCC89], a benchmark Bayesian network with 37 attributes implementing an alarm message system for patient monitoring. We sample datasets[21] for each size $n \in \{100, 500, 1000, 2000, 5000, 10000, 20000\}$, and evaluate the performance in terms of precision, recall, and F1, averaging across all 37 attributes as targets. For more accurate results, we sample 10 datasets per n and average.

First, we consider a comparison against test-based Markov blanket discovery algorithms, and more specifically, the data-inefficient baseline IAMB[22] family [TASS03], and the data-efficient state-of-the-art HITON [ATS03] and MMMB [TAS03]. We use the Causal Explorer software [ASTB03] for the aforementioned algorithms

[21] The network configurations can be found here: `https://www.bnlearn.com/bnrepository/`.

[22] We consider IAMB, IAMBPC that uses the PC algorithm [SGS93, Sec. 5.4.2] for shrinking, inIAMB that interleaves the grow-shrink phase to keep the size of the conditioning set as small as possible, and inIAMBPC.

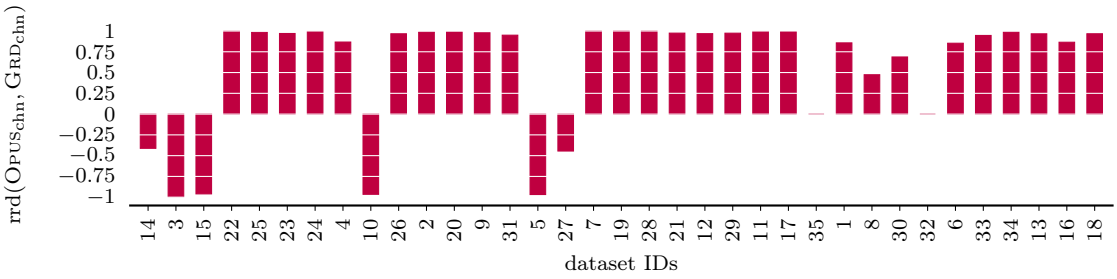


Figure 3.10: Comparing exact and heuristic algorithms in terms of running time. Relative time difference for GRD_{chn} and OPUS_{chn}. Positive (negative) numbers indicate that GRD_{chn} (OPUS_{chn}) is proportionally more effective. Datasets are sorted in decreasing number of attributes.

with the default settings ($\alpha = 0.05$ for tests). Since we know this benchmark dataset satisfies the faithfulness assumption, we use for score-based algorithm the greedy combined with shrinking ($\phi = 0.01$) with the permutation (conditional) fraction of information $\hat{F}_0$. Recall that the shrink step is necessary to guarantee the discovery of the unique Markov blanket. We denote this approach with $\text{SHRK}_{\hat{F}_0}$. We show the results in Figure 3.11. We observe that the performance of greedy with shrinking is on par with the test-based approaches. Regarding the data efficiency of the test-based approaches, the data-efficient HITON and MMMB are only marginally better than IAMB in terms of F1. So while superior in theory, in practice HITON and MMMB restrict the size of the conditioning set and hence they become approximate algorithms. Lastly, note that the test-based approaches and Causal Explorer are highly specialized towards Markov blanket discovery implementing various "tricks", e.g., interleaving, while the score-based approach we consider is a simple greedy optimization algorithm— the performance can be improved using similar techniques.

Next, we evaluate the different fraction of information estimators combined with the greedy algorithm in Figure 3.12. Note that here we do not use the shrink step as not all estimators have conditional fraction of information estimators proposed. We observe that the permutation estimator $\hat{F}_0$ clearly outperforms all other estimators on this task. The low precision behavior is attributed to positive biases (e.g., for plugin $\hat{F}_{\text{pl}}$) meaning that a lot of false positives enter the solution,

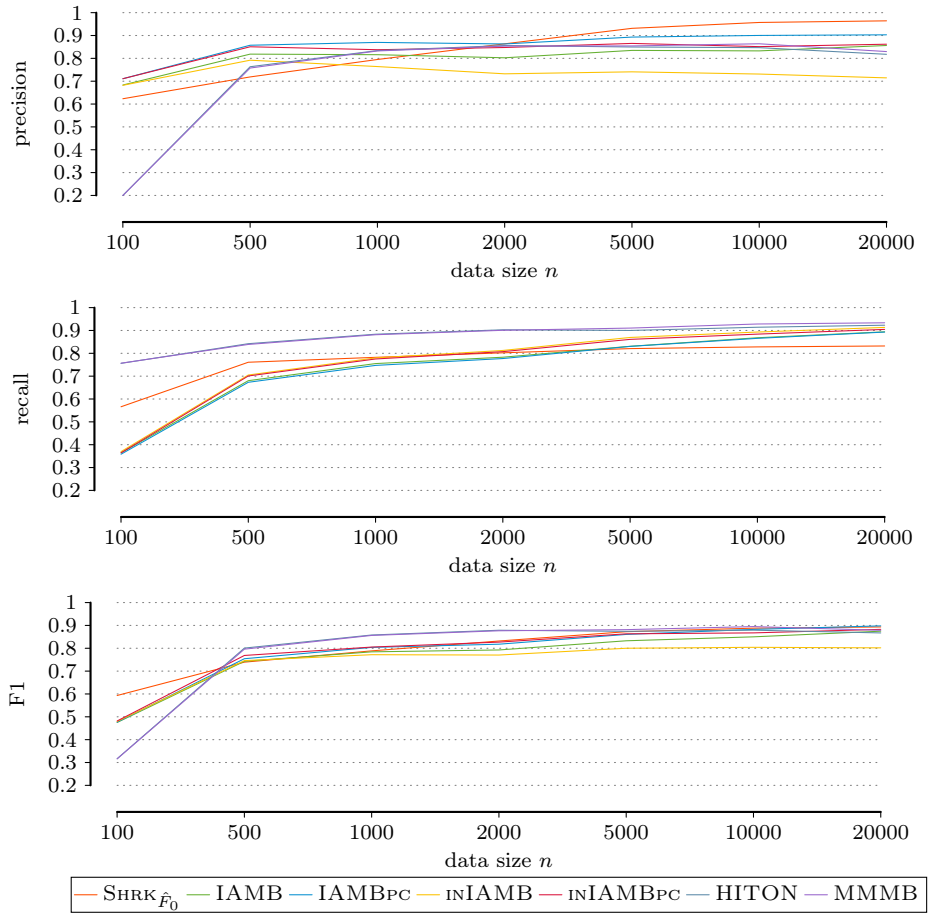


Figure 3.11: Precision, recall, and F1 of score-based versus test-based algorithms for Markov blanket discovery on the Alarm dataset. Evaluating the greedy algorithm (Alg. 3) and the permutation fraction of information with shrinking step (Alg. 4 with $\phi = 0.01$) for Markov blanket discovery versus algorithms that employ conditional independence tests.

as well as the large negative biases that contribute to a large error.

3.5.4 Case studies

We close this section with examples of concrete dependencies discovered in two different applications: determining the winner of a Tic-tac-toe board configuration and predicting the preferred crystal structure of octet binary semiconductors. Both

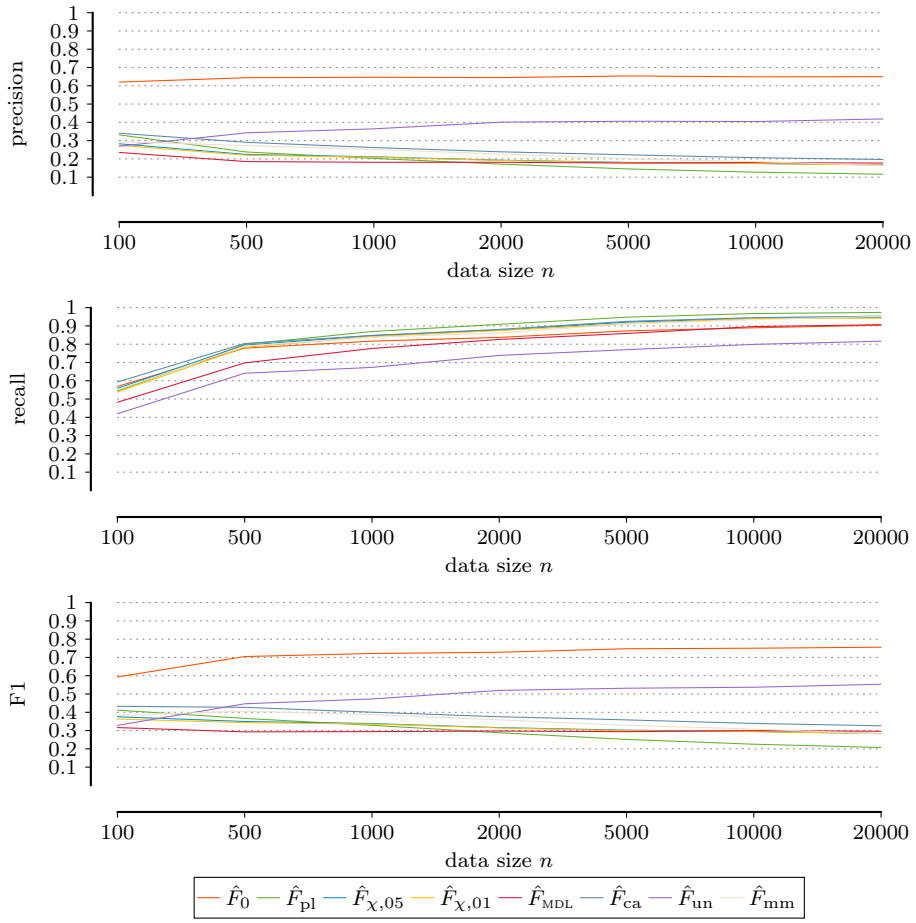


Figure 3.12: Precision, recall, and F1 of different estimators combined with the greedy algorithm for Markov blanket discovery on the Alarm dataset.

settings are examples of problems where elementary input features are available, but to correctly represent the input/output relation either non-linear models have to be used or—if interpretable models are sought—complex auxiliary features have to be constructed from the given elementary features.

Tic-tac-toe. The game of Tic-tac-toe [MR89] is one of the earliest examples of this complex feature construction problem. There are two players where each player picks a symbol from $\{x, o\}$ and, taking turns, marks his symbol in an unoccupied cell of a 3×3 game board. A player wins the game if he marks 3 consecutive cells in a row, column, or diagonal. A game can end in draw, if the

$\boldsymbol{X_1}$	X_2	$\boldsymbol{X_3}$
X_4	$\boldsymbol{X_5}$	X_6
$\boldsymbol{X_7}$	X_8	$\boldsymbol{X_9}$

3	2	3
2	4	2
3	2	3

Figure 3.13: Top-1 dependency on Tic-tac-toe with win/loss Y as target variable. Left: board with input variables in corresponding board positions, and variables contained in top dependency marked in red. **Right:** number of winning combinations each position is involved in.

$\boldsymbol{X_1}$	X_2	X_3
X_4	$\boldsymbol{X_5}$	X_6
X_7	X_8	$\boldsymbol{X_9}$

(a) top-1

X_1	X_2	$\boldsymbol{X_3}$
X_4	$\boldsymbol{X_5}$	X_6
$\boldsymbol{X_7}$	X_8	X_9

(b) top-2

$\boldsymbol{X_1}$	X_2	$\boldsymbol{X_3}$
X_4	$\boldsymbol{X_5}$	X_6
$\boldsymbol{X_7}$	X_8	$\boldsymbol{X_9}$

(c) top-3

Figure 3.14: Top-3 dependencies on Tic-tac-toe with X_5 as target variable. Target X_5 is in blue, while red indicates part of the solution. Note that all three include the win/loss attribute that is not shown here as part of the solution.

board configuration does not allow for any winning move. The dataset consists of 958 end game winning configurations (i.e., there are no draws). The 9 input variables $\mathcal{I} = \{X_1, \ldots, X_9\}$ represent the cells of the board, and can have 3 values $\{x, o, b\}$, where b denotes an empty cell (see Fig. 3.13). The target variable Y with $V_Y = \{\text{win}, \text{loss}\}$ is the outcome of the game for player x.

Searching for the top-1 dependency $\mathcal{X}^* \subseteq \mathcal{I}$ reveals as pattern with empirical fraction of information $\hat{F}_{\text{pl}}(\mathcal{X}^*; Y) = 0.61$ and corrected score $\hat{F}_0(\mathcal{X}^*; Y) = 0.45$ the variable set $\mathcal{X}^* = \{X_1, X_3, X_5, X_7, X_9\}$ i.e., the four corner cells and the middle one, which we show in Figure 3.13. This is a sensible discovery as these cells correspond exactly to those involved in the highest number of winning combinations. Removing a variable results in the loss of a considerable amount of information, while adding a variable would provide more information, but also redundancy.

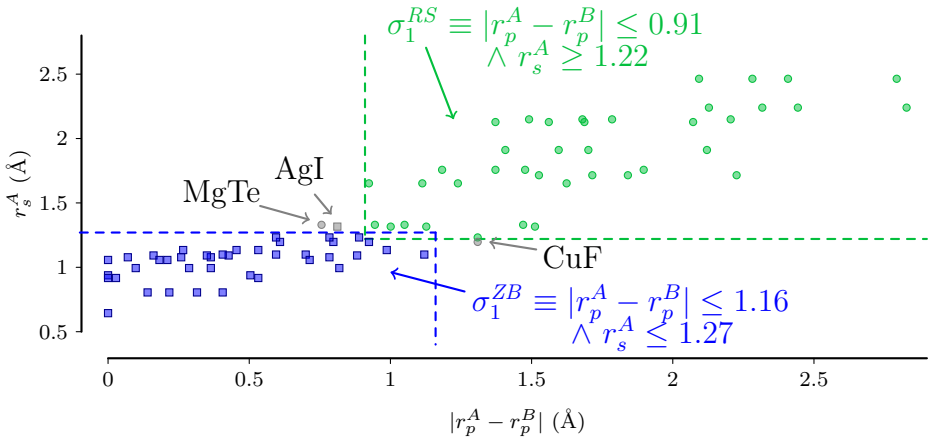


Figure 3.15: Materials Science example. Binary semiconductors that crystalize as zinkblende (boxes) and rocksalt (circles). Blue and green materials are correctly classified by subgroup-based prediction model—the involved rules (annotated) use elements of the top dependency discovered. (source: [GBV+17])

That is, the increase of fraction of information would not be higher than the increase of b_0. For top-k, the next results have also cardinality of 5, but with corrected score 0.37. Out of the MB discovery algorithms of Section 3.5.3, only HITON and MMMB were able to discover the same result, with the rest finding a solution of size 4.

In Figure 3.14 we plot the top-3 results with the target variable now being X_5, while the win/loss attribute is part of the input variables. Note that all three results include win/loss Y as part of the solution. Again the discoveries are meaningful, with top-1 and 2 having the same score 0.29, and top-3 having 0.27. The MB discovery algorithms recover either the top-1 or top-2 as a solution, since they can only identify one MB.

Materials Science. Our second example is a classical problem from Materials Science [VV69], which has meanwhile become a canonical example for the challenge of the automatic discovery of interpretable and physically meaningful prediction models of material properties [GVL+15, GBV+17]. The task is to predict the symmetry or crystal structure in which a given binary compound semi-conductor material will crystalize. That is, each of the 82 materials involved consist of

two atom types (A and B) and the output variable $Y = \{\text{rocksalt}, \text{zincblende}\}$ describes the crystal structure it prefers energy-wise. The input variables are 14 electro-chemical features of the two atom types considered in isolation: the radii of the three different electron orbitals shapes s, p, and d of atom type A denoted as $r_s(A), r_p(A), r_d(A)$, as well as four important energy quantities that determine its chemical properties (electron affinity, ionization potential, HOMO and LUMO energy levels); the same variables are defined for component B.

For this dataset, the top dependencies of cardinality 2 and identical $\hat{F}_0$ score of 0.702 are combinations of $\{r_s(A), r_p(A), r_p(B), EA(B)\}$, i.e., the atomical s and p radii of component A, the p radii of component B, and the electronic affinity of B. Again, this is a sensible finding, since three of them are contained in the best structure prediction model that can be identified using the non-linear subgroup discovery approach of Goldsmith et al. [GBV+17] (see Fig. 3.15). Also, all features are parts of the best linear LASSO model based on systematically constructed non-linear combinations of the elementary input variables by Ghiringhelli et al. [GVL+15]. The fact that not all variables of those models are identified can likely be explained by the facts that the continuous input variables are discretized and the dataset is extremely small with only 82 entries, which renders the discovery of reliable patterns with more than two variables very challenging.

The MB discovery algorithms of Section 3.5.3 fail in this task: the MMMB and HITON output a solution of size 9, i.e., almost all atomic properties from both materials A and B, while the IAMB family and GS output a solution of size 1.

Table 3.5: Datasets from Section 3.5.2, with number of rows, columns, and classes. The α values are the maximum guarantee in increments of 0.05 for OPUS to finish in ≤ 90 minutes. The $\hat{F}_0$ columns are the quality of the solutions for exhaustive and greedy search. Maximum depth is the maximum search level for OPUS.

						time(s)				$\hat{F}_0$		depth	
ID	dataset	#rows	#attr.	#cl.	α	OPUS_{chn}	OPUS_{mon}	GRD_{chn}	GRD	OPUS	GRD	max	sol.
1	*australian*	690	14	2	1.00	7.0	8.3	1.0	1.0	0.54	0.54	8	4
2	*chess*	3196	36	2	0.75	192.1	545.9	2.5	3.6	0.77	0.87	5	5
3	*coil2000*	9822	85	2	0.05	1.0	1.0	189.1	294.4	0.06	0.17	1	1
4	*connect-4*	67557	42	3	0.10	1236.8	951.5	164.3	174.8	0.10	0.29	6	4
5	*fars*	100968	29	8	0.65	3.0	7.0	93.9	119.8	0.66	0.68	2	2
6	*flare*	1066	11	6	1.00	6.8	3.2	1.0	1.0	0.65	0.65	10	3
7	*german*	1000	20	2	1.00	931.5	960.1	1.0	1.0	0.21	0.21	11	6
8	*heart*	270	13	2	1.00	1.9	1.9	1.0	1.0	0.42	0.42	7	4
9	*ionosphere*	351	33	2	1.00	46.4	47.6	1.0	1.0	0.62	0.58	5	3
10	*kddcup*	494020	41	23	0.90	18.1	37.8	520.2	616.4	0.97	0.99	2	2
11	*letter*	20000	16	26	1.00	659.5	1501.0	3.8	19.1	0.60	0.60	6	5
12	*lymph.*	148	18	4	1.00	31.2	20.2	1.0	1.0	0.48	0.45	10	5
13	*magic*	19020	10	2	1.00	38.5	31.6	1.3	1.3	0.43	0.43	8	5
14	*move-libras*	360	90	15	0.50	1.0	266.6	1.7	25.9	0.32	0.32	3	2
15	*optdigits*	5620	64	10	0.35	1.0	4.3	25.1	139.3	0.36	0.53	2	2
16	*pageblocks*	5472	10	5	1.00	7.4	5.2	1.0	1.0	0.65	0.60	8	4
17	*penbased*	10992	16	10	1.00	233.6	277.5	1.6	5.6	0.75	0.75	7	4
18	*poker*	1025010	10	10	1.00	2594.7	1705.2	86.0	205.3	0.57	0.57	7	5
19	*ring*	7400	20	2	1.00	1393.9	1197.3	1.1	7.4	0.29	0.29	6	4
20	*satimage*	6435	36	7	0.80	173.8	954.4	2.0	27.6	0.74	0.74	4	4

21	*segment*	2310	19	7	1.00	39.1	53.3	1.0	1.2	0.84	0.84	9	3
22	*sonar*	208	60	2	1.00	403.5	431.9	1.0	3.8	0.34	0.32	5	3
23	*spambase*	4597	57	2	0.55	515.6	574.6	15.4	50.1	0.54	0.60	7	4
24	*spectfheart*	267	44	2	1.00	171.1	369.3	1.0	1.9	0.23	0.22	5	3
25	*splice*	3190	60	3	0.65	92.3	851.0	1.5	46.9	0.65	0.65	4	4
26	*texture*	5500	40	11	0.80	62.9	480.3	2.1	36.6	0.76	0.76	5	4
27	*thyroid*	7200	21	3	0.50	1.0	5.8	1.8	2.0	0.50	0.50	3	3
28	*twonorm*	7400	20	2	1.00	1332.2	1162.4	1.3	7.4	0.42	0.42	6	4
29	*vehicle*	846	18	4	1.00	38.2	53.2	1.0	1.0	0.48	0.48	8	3
30	*vowel*	990	13	11	1.00	3.2	5.1	1.0	1.0	0.45	0.45	5	3
31	*wdbc*	569	30	2	1.00	19.9	38.2	1.0	1.2	0.76	0.75	7	3
32	*wine*	178	13	3	1.00	1.0	1.0	1.0	1.0	0.71	0.71	3	2
33	*wine-red*	1599	11	11	1.00	18.7	10.3	1.0	1.0	0.20	0.20	7	3
34	*wine-white*	4898	11	11	1.00	77.4	36.2	1.0	1.0	0.19	0.19	8	5
35	*zoo*	101	15	7	1.00	1.0	4.1	1.0	1.0	0.80	0.75	7	5
avg.		52000	30	6.4	0.85	296	360	32	51	0.51	0.53	5.9	3.6

3.6 Discussion and conclusions

We considered the dual problem of measuring and efficiently discovering functional dependencies from data, and for effective knowledge discovery, we investigated the combinatorial optimization problem of maximizing the fraction of information F. This problem is theoretical justified and the results have causal interpretations under standard assumptions. To overcome the bias arising from high-dimensional distributions and correct the inflated estimates, we proposed a consistent and robust estimator for mutual information based on the expected value of the null distribution. Concerning the optimization problem, we proved NP-hardness and derived two bounding functions for the estimator that can be used to prune the search space. With these, we can effectively discover the optimal, or α-approximate top-k dependencies with branch-and-bound. The experimental evaluation showed that the estimator has desired statistical properties, the bounding functions are very effective with both exhaustive and heuristic algorithms, and the greedy algorithm provides solutions that are nearly optimal. Assuming a Bayesian network, our resulting method for Markov blanket discovery is on par with the state-of-the-art algorithms based on independence tests. Qualitative experiments on two case studies indicate that our proposed framework indeed discovers informative dependencies corroborated by domain experts, while independence testing Markov blanket discovery algorithms under-perform.

3.6.1 Greedy optimization

While the given reduction from set cover can be extended to show that, unless P=NP, no fully polynomial time approximation scheme exists, the possibility for weaker approximation guarantees remains. In particular, the strong empirical performance of the greedy algorithm hints that $\hat{F}_0$ could have a certain structure favored by the greedy algorithm, e.g., some weaker form of submodularity (we remind that $\hat{F}_0$ is neither submodular nor monotone). For instance, one could explore ideas from Horel and Singer [HS16] where a monotone function is ϵ-approximately submodular if it can be bounded by a submodular function within $1 \pm \epsilon$. Another idea is that of restricted submodularity for monotone functions [DGP+08], where a

function is submodular over a subset of the search space. Perhaps the most promising is the submodularity index for general set functions [ZS16], where a proxy for the degree of non-submodularity is incorporated in the approximation guarantee.

3.6.2 Significance testing and multiple hypotheses

An important aspect for further investigation is connections of our correction approach to multiple hypothesis testing. This is a topic in statistics concerned with problems arising from performing significance tests on a large volume of hypotheses (see Hämäläinen and Webb for a great tutorial [HW19]). Let us consider the following example [HW19, Sec. 6]. We perform tests on $m \in \mathbb{Z}^+$ true null hypotheses at significance level α. Assuming that the probability of type I error is exactly α, then we should be expecting $m \cdot \alpha$ false positives on average. For example, with $m = 100000$, we can expect 5000 false positives. Such numbers are common in typical data analysis tasks. To overcome this problem, methods try to control the familywise error rate (FWER) or the false discovery rate (FDR) by making the rejection of null hypotheses harder. A basic approach for this is the Bonferroni correction that uses an adjusted significance level $\alpha' = \alpha/b$, where b is the total number of hypotheses.

Our approach adjusts the estimation such that it is unbiased under the null hypothesis. That is, we aim to reduce the number of false positives by correcting the estimation error when the variables are independent. For a fixed number of samples n, both the plugin estimator and the correction term monotonically increase with the superset relation. In general, they both have the tendency to increase with the domain size. If the domain sizes are large compared to n, the estimates are penalized heavier. This is very similar to the approach of Webb [Web08], where the α values are adjusted according to the search level, with higher levels being penalized more.

Some directions to investigate are connections of our approach to significance testing and multiple hypotheses, e.g., implications of subtracting the mean of the null distribution. Is it related to performing tests at level $\alpha = 0.5$? Since we are performing a top-k formulation, does it mean we stop searching for larger

candidates the moment such a test fails? Can we argue about FWER and FDR as a function of k? It would be also interesting to argue further about the benefits of using this correction approach by relating it to multiple tests for exploratory research [GS11], where in a nutshell, it is more favorable to retrieve a larger number of potential interesting hypotheses, instead of using strict tests to reduce false positives.

3.6.3 Markov blanket discovery

Assuming a Bayesian network, maximizing the fraction of information can be seen as a dual formulation to independence testing for Markov blanket discovery. While the evaluation showed that our score-based approach is competitive on the Alarm network, it is important to note the main advantage of the testing approach: the data efficient algorithms, e.g., HITON, MMMB, are better-suited for recovering larger Markov blankets. For example, if the Bayesian network has an MB of size 60, it is unreasonable to expect our approach, GS, and IAMB, to retrieve it, unless the number of samples is enormous. However, HITON and MMMB are approximate, as in order to be computationally efficient, they consider a maximum size for the conditional set. That is, these methods conclude that a variable should enter the Markov blanket not by testing all possible subsets of the currently selected MB, but only subsets up to a certain size. Note that as the sizes of Markov blankets increase, these methods have to become more and more approximate to maintain efficiency. To summarize, we conclude that exact methods, i.e., methods that consider jointly all the variables, are statistically inefficient, while the data efficient are computationally infeasible and approximate. For very large Markov blankets, standard feature selection, e.g., MRMR, should be preferred, but note that they do not guarantee the discovery of a Markov blanket.

The advantages of score-based approaches on the other hand are two. First, searching for the exact and maximal top-k corresponds to identifying multiple Markov blankets. It would be interesting to see a comparison with methods for multiple Markov blankets such as KIAMB [PNBT07], EGSG [LLZ10], TIE [SLA13], and also derive criteria for identifying the number k of possible Markov blan-

kets. Second, since it is a combinatorial maximization problem, it allows for approximate algorithms such as greedy, accelerated greedy [Min78], and stochastic greedy [MBK+15], that can scale to problems of arbitrary size. It would be interesting to see under what conditions, e.g., faithfulness, there can be approximation guarantees.

3.6.4 Future work

Perhaps the most interesting direction is that of efficiently arriving at a diverse set of top-k solutions. Here, we used a top-k formulation without considering maximal solutions or any post-processing. Indeed, such formulations would come closer to the objective of Theorem 3.0.1 for discovering the k Markov blankets. Regarding efficiency, Pennerath [Pen18] introduce efficient algorithms to compute entropic measures for large k based on FP-Growth. This framework could potentially be extended to retrieve maximal solutions, e.g., with the result set represented by a prefix tree, and only reporting root-to-leaf paths. Moreover, Pennerath [Pen10] also introduce efficient post-processing techniques to retrieve diverse sets from top-k with the notion of locally optimal patterns.

4

Discovering robust totally correlated sets

In this chapter we consider categorical input variables $\mathcal{I} = \{X_1, \ldots, X_d\}$ and propose algorithms for discovering subsets $\mathcal{X} \subseteq \mathcal{I}$ that exhibit high mutual dependency and can summarize aspects of the process under consideration.

For our knowledge discovery purposes, existing solutions for this problem have several drawbacks. Many methods are primarily defined for binary data and measure only pairwise associations with interestingess functions such as chi-square statistic [BMS97], all-confidence [Omi03], h-confidence [XTK06], or mutual information [KCN08]. By considering only pairwise associations, higher-order interactions among the features are neglected. In addition, data transformations from categorical attributes to boolean may incur information loss. Finally, such methods are parameterized with various thresholds, e.g., minimum all-confidence, leading to an uncontrollable output size, i.e., they might miss interesting dependencies or receive too many. In a nutshell, we find that unsupervised mining methods, although relevant for their own respective applications, lack a comprehensive formalization of dependency, as well as parameter-free, single-objective optimization problems for categorical data that we are interested in.

This chapter is an extended version of work that originally appeared in IEEE International Conference on Data Mining (ICDM) [MBV19].

In this chapter, we build upon the concept of **total correlation**[1] $W(\mathcal{X})$ for sets $\mathcal{X} \subseteq \mathcal{I}$, the multivariate extension of mutual information, which quantifies the amount of shared information in a set of random variables while being agnostic about the type of relationship [Wat60]. Total correlation has been successfully employed in other unsupervised scenarios, such as learning latent representations [SG14], measuring correlation in real-valued data [NMV16, WRN+17], and mining high order interactions in binary data [ZPWN08]. Without appropriate normalization, however, scores over sets of different cardinalities are not comparable, which is a problem when searching for the top dependent subsets $\mathcal{X}^* \subseteq \mathcal{I}$. We hence consider **normalized total correlation** $w(\mathcal{X})$, which does not only address this, but is also interpretable: a score of 0 means the random variables in a set are statistically independent, and a score of 1 that there exists a variable that "explains" all others. With this, we are then looking solutions to the combinatorial optimization problem of finding the top-k subsets $\mathcal{X}_1^*, \ldots, \mathcal{X}_k^* \subseteq \mathcal{I}$ with

$$w(\mathcal{X}_i^*) = \max\{w(\mathcal{X}) \colon\ w(\mathcal{X}_{i-1}^*) \geq w(\mathcal{X}), \mathcal{X} \subseteq \mathcal{I}\}\ . \tag{4.1}$$

Although theoretically sound, in practice normalized total correlation suffers from inflated estimates when computed from empirical data. This is not a surprise since it is an extension of mutual information. In fact, compared to the supervised scenario of the previous chapters, the situation here is more "chaotic" since total correlation is a sum of increasingly higher-dimensional mutual information terms (see Fig. 4.2 for a demonstration). As for the resulting combinatorial optimization problem Eq.(4.1), putting aside the exponential search space, the lack of a special attribute (i.e., a target Y) that can lead to a more informed search, as well as the existence of a normalizer that is not constant throughout the search process, make it harder to obtain admissible bounding functions for pruning.

To solve these, we build upon the previous chapter and propose a robust and efficient estimator for normalized total correlation. Furthermore, we enable effective exact and heuristic algorithms for the discovery of the top dependent sets

[1] Correlation here does not imply linear relationships. In that sense, total dependency would be more accurate.

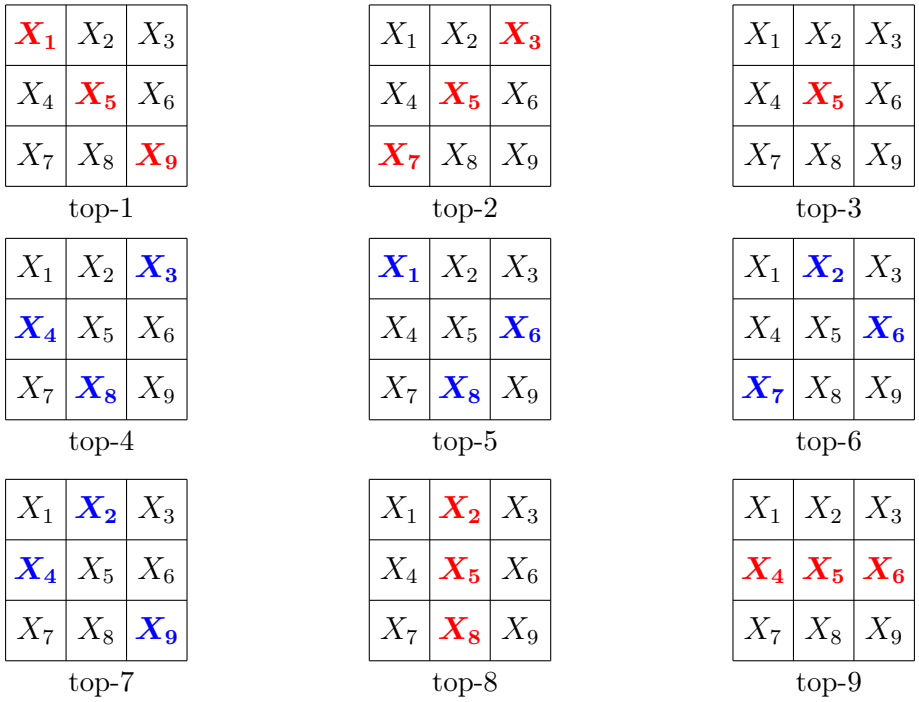


Figure 4.1: Top-9 dependent sets discovered on Tic-tac-toe with our proposed solution. Color indicates the selected cells, with red designating the inclusion of X_{10} that corresponds to the binary outcome of the game. In a nutshell, red and blue dependent sets can be interpreted as descriptions for win and loss, respectively. (Section 4.4.3)

by exploiting various structural properties of the estimator proposed. Our main contributions are the following:

- we propose a consistent, robust, and efficient estimator for the normalized total correlation (Sec. 4.2),
- we derive admissible bounding functions for effective pruning and provide algorithms for exact, approximate, and heuristic search (Sec. 4.3), and finally,
- we perform evaluation on a wide range of real and synthetic datasets (Sec. 4.4, see Fig. 4.1 for a demonstration).

We introduce the normalized total correlation in Section 4.1 and round up with discussion and conclusions in Section 4.5.

4.1 Normalized total correlation

Introduced by Watanabe [Wat60], the **total correlation** for set of variables $\mathcal{X} = \{X_1, \ldots, X_m\}$ with joint probability distribution $p(\mathcal{X})$ 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. Essentially, total correlation is a multivariate dependency/redundancy measure quantifying the total amount of shared information in a set of random variables. It holds that $W(\mathcal{X}) \geq 0$, with equality if and only if all variables $X \in \mathcal{X}$ are statistically independent, and is monotonically increasing with the subset relation, i.e., for sets of variables $\mathcal{X}$ and $\mathcal{X}'$ with $\mathcal{X} \subseteq \mathcal{X}'$, it holds that $W(\mathcal{X}) \leq W(\mathcal{X}')$. Note that total correlation can be expressed as the KL-divergence between the joint $p(\mathcal{X})$ and the product of marginals $\prod_{X \in \mathcal{X}} p(X)$ and that it is order invariant as a functional of p.

Total correlation, however, is not suitable for comparing the degree of dependency between different sets of variables, since set cardinalities, as well as the joint and marginal entropies of the variables involved, all can vary. In addition, the monotonicity property implies that larger sets are more preferable as solutions, even in situations where $W(\mathcal{X}') = W(\mathcal{X}) + \epsilon$ for sets $\mathcal{X} \subseteq \mathcal{X}'$. This introduces redundancy and might hinder next steps of the analysis, such as visualizations. Finally, total correlation lacks an intuitive and intepretable scale, e.g., in $[0, 1]$, that would facilitate the process to understand the results and reason about. 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.

Proposition 4.1.1. *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.,* $X_j = f(X_i), \forall X_j \in \mathcal{X}$.

Proof. a) We upper-bound $W(\mathcal{X})$ by lower bounding $H(\mathcal{X})$. Since Shannon entropy is monotonically increasing with the subset relation, we have that $H(\mathcal{X}) \geq \max_{X\in\mathcal{X}} H(X)$, and hence $W(\mathcal{X}) \leq \sum_{X\in\mathcal{X}} H(X) - \max_{X\in\mathcal{X}} H(X)$.

b) Suppose that $W(\mathcal{X}) = \sum_{X\in\mathcal{X}} H(X) - \max_{X\in\mathcal{X}} H(X)$. Then $H(\mathcal{X}) = \max_{X\in\mathcal{X}} H(X) = H(X_q)$ for some $q \in [1, m]$. Using the chain rule for entropy, i.e., $H(\mathcal{X}) = \sum_{i=1}^{m} H(X_i \,|\, \mathcal{X}_{i-1})$, and since this decomposition is order-invariant, it is clear that $H(X_i \,|\, X_q) = 0$ for all $X_i \in \mathcal{X}$. This is possible if and only if $X_i = f(X_q)$ for all $X_i \in \mathcal{X}$.

Conversely, suppose there exists $X_q \in \mathcal{X}$ s.t., $X_j = f(X_q), \forall X_j \in \mathcal{X}$. Hence, we have that $H(X_j \,|\, X_q) = 0$ for all $X_j \in \mathcal{X}$, and $H(\mathcal{X}) = H(X_q)$. Now, $X_q = \max_{X\in\mathcal{X}} H(X)$, i.e., X_q must be the variable with the highest entropy, hence $W(\mathcal{X}) = \sum_{X\in\mathcal{X}} H(X) - \max_{X\in\mathcal{X}} H(X)$. □

We now define the **total correlation upper-bound** as $\bar{W}(\mathcal{X}) = \sum_{X\in\mathcal{X}} H(X) - \max_{X\in\mathcal{X}} H(X)$, and proceed to define 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 other.[2] 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.

Now given empirical data $\mathbf{D}_n$, estimating the information-theoretic quantities involved in w poses the same problem as in the previous chapters, i.e., inflated estimates. Here the problem is in fact more profound: while it is easier in general to obtain good estimates for marginal quantities, e.g., the normalizer of w, total correlation involves mutual information terms that need to be estimated for increasingly larger sets of variables. This can lead to situations with arbitrary estimates (see Fig. 4.2 for a demonstration).

[2]Note that the bound for total correlation is in general known in the literature, e.g., [Wat60]. However, a formal proof for the bound is often missing, which we present here for both self-containment, and to better understand its properties.

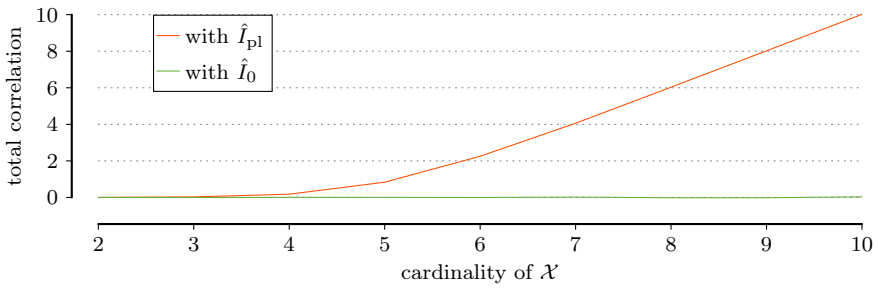


Figure 4.2: Inflated dependency. Estimated total correlation for variable set $\mathcal{X}$ of increasing cardinality. All variables are uniformly and independently sampled with domain size 4 and sample size 1000. Population value for total correlation is 0. Dependency increases when plugin estimator $\hat{I}_{\text{pl}}$ is used, but not for the permutation $\hat{I}_0$.

4.2 Permutation normalized total correlation

In this section we derive a robust, consistent, and efficient to compute estimator for the normalized total correlation. Following the same correction principle as in Chapter 3, and assuming we can adequately estimate marginal entropies $\hat{H}_{\text{pl}}(X)$, we can define a robust estimator for the normalized total correlation by plugging $\hat{I}_0$ and arrive at

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

However, unlike the plugin $\hat{w}_{\text{pl}}$, 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.,

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

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

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.

The upper-bound we consider is the one by Nguyen et al. [NEB10, Thm. 7] that we also introduced in Section 3.2, i.e., for variables X, Y, with domain sizes S_X, S_Y, and sample size n, it holds that $m_0(X, Y, n) \leq \log \frac{n+S_X S_Y - S_X - S_Y}{n-1}$. We denote this **upper-bound** with $m_{\bar{0}}(X, Y, n)$, and the corresponding **correction term** with $t_{\bar{0}}(\mathcal{X}, n)$, 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 $\mathbf{x} \in V_{\mathcal{X}}$, then 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 **relaxed upper-bound** $m_{\bar{\bar{0}}}$ with

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

and to the following **relaxed correction term** $t_{\bar{\emptyset}}$ with

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

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

Theorem 4.2.1. *For set of variables $\mathcal{X} = \{X_1, \ldots, X_m\}$, it holds*

a) $t_{\bar{\emptyset}}(\mathcal{X}, n) \geq t_{\emptyset}(\mathcal{X}, n)$

b) $\lim_{n\to\infty} t_{\bar{\emptyset}}(\mathcal{X}, n) = 0$

c) $\sum_{i=2}^{m} m_{\bar{\emptyset}}(\mathcal{X}_{\sigma(i-1)}, X_{\sigma(i)}, n)$ *is maximized for* $\sigma^* \in S_m$ *with* $S_{X_{\sigma^*(1)}} \geq S_{X_{\sigma^*(2)}} \cdots \geq S_{X_{\sigma^*(m)}}$

Proof. For readability, we drop σ as a subscript whenever clear from the context. We prove (a) by first showing that it holds for any $\sigma \in S_m$. Given a $\sigma \in S_m$, and any $i \in [2, m]$, we have

$$\begin{aligned} m_{\bar{\emptyset}}(\mathcal{X}_{i-1}, X_i, n) =& \log \frac{n + S_{X_i} \prod_{X \in \mathcal{X}_{i-1}} S_X}{n-1} \\ \geq& \log \frac{n + S_{X_i} \prod_{X \in \mathcal{X}_{i-1}} S_X - \prod_{X \in \mathcal{X}_{i-1}} S_X - S_{X_i}}{n-1} \\ =& m_{\emptyset}(\mathcal{X}_{i-1}, X_i, n) \ . \end{aligned}$$

Since this holds for any $\sigma \in S_m$ and $i \in [2, m]$, then for the σ^* with $\sigma^* = \arg\max_{\sigma \in S_m} \sum_{i=2}^{m} m_{\emptyset}(\mathcal{X}_{\sigma(i-1)}, X_{\sigma(i)}, n)$ we have that $\sum_{i=2}^{m} m_{\bar{\emptyset}}(\mathcal{X}_{\sigma^*(i-1)}, X_{\sigma^*(i)}, n)$ is larger. Statement (b) follows from $\lim_{n\to\infty} \log(\frac{(n+a)}{(n-1)}) = 0$.

For (c) let us consider a $\sigma^* \in S_m$ for which $S_{X_{\sigma^*(1)}} \geq \cdots \geq S_{X_{\sigma^*(m)}}$, and any arbitrary $\sigma \in S_m$. We prove this statement by doing a pairwise comparison between

$m_{\bar{\emptyset}}(\mathcal{X}_{\sigma(i-1)}, X_{\sigma(i)}, n)$ and $m_{\bar{\emptyset}}(\mathcal{X}_{\sigma^*(i-1)}, X_{\sigma^*(i)}, n)$ for any $i \in [2, m]$. We have

$$\begin{aligned} m_{\bar{\emptyset}}(\mathcal{X}_{\sigma^*(i-1)}, X_{\sigma^*(i)}, n) =& \log \frac{n + \prod_{X \in \mathcal{X}_{\sigma^*(i)}} S_X}{n-1} \\ \geq& \log \frac{n + \prod_{X \in \mathcal{X}_{\sigma(i)}} S_X}{n-1} \\ =& m_{\bar{\emptyset}}(\mathcal{X}_{\sigma(i-1)}, X_{\sigma(i)}, n) \ , \end{aligned}$$

where the inequality follows from the fact that $\prod_{X \in \mathcal{X}_{\sigma^*(i)}} S_X$ is the product of the i largest domain sizes. Since this holds for any $\sigma \in S_m$ and $i \in [2, m]$, then $\sigma^* = \arg\max_{\sigma \in S_m} \sum_{i=2}^{m} m_{\bar{\emptyset}}(\mathcal{X}_{\sigma(i-1)}, X_{\sigma(i)}, n)$. □

We now have an efficiently computable correction term $t_{\bar{\emptyset}}(\mathcal{X}, n)$, 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 $t_{\bar{\emptyset}}$, this correction is as conservative with regards to its estimates, which is a design goal for robustness. Finally, we arrive at the **permutation normalized total correlation**[3]

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

In addition to being very efficient, the consistency of the plugin $\hat{H}_{\mathrm{pl}}$ [AK01], together with Theorem 4.2.1b), implies that $\hat{w}_{\bar{\emptyset}}$ is a consistent estimator for the normalized total correlation. The estimators discussed here are evaluated further for their statistical properties in Section 4.4.1.

4.3 Optimization algorithms

Here, we provide algorithms for the optimization problem of Eq.(4.1). Given the combinatorial nature of the problem, as well as the hardness result for optimizing the permutation mutual information (Sec. 3.2), it is unlikely that the optimization of $\hat{w}_{\bar{\emptyset}}$ allows for a polynomial algorithm. While the complexity of the optimization

[3]Note that upper-bounded should have been part of the name, but for simplicity we omit it.

problem under consideration is an open question, here we derive two practically efficient algorithms for exact and heuristic search. For both, we derive admissible bounding functions for $\hat{w}_{\bar{\emptyset}}$ to be used for pruning.

Recall that an admissible bounding function (see Sec. 3.3) is an upper-bound for the maximum attainable score $\hat{w}_{\bar{\emptyset}}(\mathcal{X}')$ for supersets of $\mathcal{X}$ in the enumerated search space. Hence, the ideal one 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 $\hat{w}_{\text{pl}}(\mathcal{X})$ and a lower bound for $t_{\bar{\emptyset}}(\mathcal{X}, n)$, in order to arrive at a looser, but efficient to compute bounding function. In our setting, 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 (i.e., plugin and correction) are monotonically increasing functions, but this property does not extend combined 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 $t_{\bar{\emptyset}}(\mathcal{X}'', n) \leq t_{\bar{\emptyset}}(\mathcal{X}, n)$. In other words, anything can happen.

As it turns out, 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 4.3.1 (**Low entropy extension**). *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 $X' \in \mathcal{X}' \setminus \mathcal{X}$, $\hat{H}_{pl}(X') \leq \min_{X \in \mathcal{X}} \hat{H}_{pl}(X)$.*

We can guarantee that this partial order holds in the enumerated search space by simply considering a **decreasing-entropy refinement operator** (see Sec. 3.3) of the form

$$r^H_{\mathcal{I}}(\mathcal{X}) = \{\mathcal{X} \cup \{X\} \colon \hat{H}_{\text{pl}}(X) \leq \min_{X' \in \mathcal{X}} \hat{H}_{\text{pl}}(X'), X \in \mathcal{I} \setminus \mathcal{X}\} \ ,$$

i.e., it holds that $\mathcal{X} \subseteq_H \mathcal{X}'$ for all $\mathcal{X}' \in r_{\mathcal{I}}^H(\mathcal{X})$. We now proceed with showing that under this partial order, the correction term $t_{\bar{\emptyset}}$ is monotonically increasing. First, we provide the following required lemma.

Lemma 4.3.1. *For two fractions a/x and b/y of positive integers, if $a/x \leq b/y$, then it holds that $a/x \leq (a+b)/(x+y)$.*

Proof. We have

$$\frac{a}{x} \leq \frac{b}{y} \Rightarrow ay \leq bx \Rightarrow ay + ax \leq bx + ax \Rightarrow$$
$$\frac{ay + ax}{x(x+y)} \leq \frac{ax + bx}{x(x+y)} \Rightarrow \frac{a(y+x)}{x(x+y)} \leq \frac{x(a+b)}{x(x+y)} \Rightarrow$$
$$\frac{a}{x} \leq \frac{a+b}{x+y} \ ,$$

concluding the proof. □

Theorem 4.3.1. *For subsets $\mathcal{X}, \mathcal{X}'$ of $\mathcal{I}$ with $\mathcal{X} \subseteq_H \mathcal{X}'$, it holds that $t_{\bar{\emptyset}}(\mathcal{X}, n) \leq t_{\bar{\emptyset}}(\mathcal{X}', n)$.*

Proof. Let $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{X}' = \mathcal{X} \cup \mathcal{Z}$, with $\mathcal{Z} = \{Z_1, \ldots, Z_q\}$. Let us assume for simplicity and w.l.o.g. that X_1 is the variable with the maximum entropy in $\mathcal{X}$, and that $S_{X_1} \geq \cdots \geq S_{X_d}$ and $S_{Z_1} \geq \cdots \geq S_{Z_q}$.[4] In addition, let us assume for now that $\min_{X \in \mathcal{X}} S_X \geq \max_{Z \in \mathcal{Z}} S_Z$.

Since $\mathcal{X} \subseteq_H \mathcal{X}'$, X_1 is also the largest entropic variable in $\mathcal{X}'$, and because $\min_{X \in \mathcal{X}} S_X \geq \max_{Z \in \mathcal{Z}} S_Z$, we can separate the contributions of $\mathcal{X}$ and $\mathcal{Z}$ and reformulate $t_{\bar{\emptyset}}(\mathcal{X}', n)$ as

$$\frac{\sum_{i=2}^m m_{\bar{\emptyset}}(\mathcal{X}_{i-1}, X_i, n) + \sum_{j=1}^q m_{\bar{\emptyset}}(\mathcal{X} \cup \mathcal{Z}_{j-1}, Z_j, n)}{\sum_{i=2}^m \hat{H}_{\mathrm{pl}}(X_i) + \sum_{j=1}^q \hat{H}_{\mathrm{pl}}(Z_j)} \ .$$

Now let us use the notation $a = \sum_{i=2}^m m_{\bar{\emptyset}}(\mathcal{X}_{i-1}, X_i, n)$, $b = \sum_{j=1}^q m_{\bar{\emptyset}}(\mathcal{X} \cup \mathcal{Z}_{j-1}, Z_j, n)$, $x = \sum_{i=2}^m \hat{H}_{\mathrm{pl}}(X_i)$, $y = \sum_{j=1}^q \hat{H}_{\mathrm{pl}}(Z_j)$. We need to show that $\frac{a+b}{x+y} \geq \frac{a}{x}$.

[4] The former allows us to write the normalizer $\bar{W}(\mathcal{X})$ as $\sum_{i=2}^m \hat{H}_{\mathrm{pl}}(X_i)$, and the latter to remove the max operator from the numerator of $t_{\bar{\emptyset}}$.

As $\mathcal{X} \subseteq_H \mathcal{X}'$, we have that $\sum_{j=1}^{q} \hat{H}_{\text{pl}}(Z_j)$ is a sum of q terms, smaller than the $m-1$ terms of $\sum_{i=2}^{m} \hat{H}_{\text{pl}}(X_i)$. In addition, and by the definition of $m_{\bar{\emptyset}}$, the quantity $\sum_{j=1}^{q} m_{\bar{\emptyset}}(\mathcal{X} \cup \mathcal{Z}_{i-1}, Z_i, n)$ is a sum of q terms larger than the $m-1$ terms of $\sum_{i=2}^{m} m_{\bar{\emptyset}}(\mathcal{X}_{i-1}, X_i, n)$. Hence, the fraction b/y is larger than that of a/x, and from Lemma 4.3.1, we have that $\frac{a+b}{x+y} \geq \frac{a}{x}$.

Now if it were not the case that $\min_{X \in \mathcal{X}} S_X \geq \max_{Z \in \mathcal{Z}} S_Z$, i.e., there exist variables in $\mathcal{Z}$ with domain sizes larger than those in $\mathcal{X}$, then we could still write the numerator of $t_{\bar{\emptyset}}(\mathcal{X}')$ as two sums a' and b' with $m-1$ and q terms respectively, and it would hold that $a' \geq a$ and $b' \geq b$, and hence

$$\frac{a}{x} \leq \frac{a+b}{x+y} \leq \frac{a'+b'}{x+y} \ ,$$

concluding the proof. □

Following from the theorem, and using the upper bound 1 for $\hat{w}_{\text{pl}}(\mathcal{X})$, we have that

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

for all $\mathcal{X}'$ that are low entropy extensions of $\mathcal{X}$, which allows us to define the **monotonicity bounding function**

$$\bar{w}_{\text{mon}}(\mathcal{X}) = 1 - t_{\bar{\emptyset}}(\mathcal{X}, n) \ . \tag{4.2}$$

It is clear, however, that Eq.(4.2) is not tight: it upper bounds $\hat{w}_{\text{pl}}(\mathcal{X})$ with the maximum possible value for the normalized total correlation, without taking into consideration both the dependency in $\mathcal{X}$ so far, nor how "good" it might actually become for $\mathcal{X}'$. We derive a much tighter upper bound for $\hat{w}_{\text{pl}}$ by further exploiting the structure of the enumerated space. We define $R_{\mathcal{X}} = \{X \colon \hat{H}_{\text{pl}}(X) \leq \min_{X' \in \mathcal{X}} \hat{H}_{\text{pl}}(X'), X \in \mathcal{I} \setminus \mathcal{X}\}$ to be the set of all refinement elements of $\mathcal{X}$ in the

enumerated search space, and $\bar{w}(\mathcal{X})$ the **branch-informed upper-bound**

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

i.e., the plugin $\hat{w}_{\text{pl}}(\mathcal{X})$ 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 $\hat{w}_{\text{pl}}(\mathcal{X})$ with respect to $\subseteq_H$.

Theorem 4.3.2. *For a $\mathcal{X} \subseteq \mathcal{I}$ and any $\mathcal{X}' \subseteq \mathcal{I}$ with $\mathcal{X} \subseteq_H \mathcal{X}'$, it holds that $\bar{w}(\mathcal{X}) \geq \hat{w}_{pl}(\mathcal{X}')$.*

Proof. Let $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{X}' = \mathcal{X} \cup \mathcal{Z}$, with $\mathcal{Z} = \{Z_1, \ldots, Z_q\}$. We have

$$\begin{aligned}
\hat{w}_{\text{pl}}(\mathcal{X}') &= \frac{\sum_{i=2}^{m} \hat{I}_{\text{pl}}(\mathcal{X}_{i-1}; X_i) + \sum_{j=1}^{q} \hat{I}_{\text{pl}}(\mathcal{X} \cup \mathcal{Z}_{j-1}; Z_j)}{\bar{W}(\mathcal{X}) + \sum_{j=1}^{q} \hat{H}_{\text{pl}}(Z_j)} \\
&\leq \frac{\sum_{i=2}^{m} \hat{I}_{\text{pl}}(\mathcal{X}_{i-1}; X_i) + \sum_{j=1}^{q} \hat{H}_{\text{pl}}(Z_j)}{\bar{W}(\mathcal{X}) + \sum_{j=1}^{q} \hat{H}_{\text{pl}}(Z_j)} \\
&\leq \frac{\sum_{i=2}^{m} \hat{I}_{\text{pl}}(\mathcal{X}_{i-1}; X_i) + \sum_{j=1}^{q} \hat{H}_{\text{pl}}(Z_j) + \sum\limits_{X' \in R_{\mathcal{X}'}} \hat{H}_{\text{pl}}(X')}{\bar{W}(\mathcal{X}) + \sum_{j=1}^{q} \hat{H}_{\text{pl}}(Z_j) + \sum_{X' \in R_{\mathcal{X}'}} \hat{H}_{\text{pl}}(X')} \\
&= \frac{\sum_{i=2}^{m} \hat{I}_{\text{pl}}(\mathcal{X}_{i-1}; X_i) + \sum_{X' \in R_{\mathcal{X}}} \hat{H}_{\text{pl}}(X')}{\bar{W}(\mathcal{X}) + \sum_{X' \in R_{\mathcal{X}}} \hat{H}_{\text{pl}}(X')} = \bar{w}(\mathcal{X}) \ ,
\end{aligned}$$

where the first inequality follows from the fact that $\hat{I}_{\text{pl}}(X;Y) \leq \min\{\hat{H}_{\text{pl}}(X), \hat{H}_{\text{pl}}(Y)\}$ for variables X and Y (Prop. 2.1.1), and that $\mathcal{X} \subseteq_H \mathcal{X}'$, i.e., $\hat{I}_{\text{pl}}(\mathcal{X} \cup \mathcal{Z}_{j-1}; Z_j) \leq \hat{H}_{\text{pl}}(Z_j)$ for all $j \in [1, q]$. The second inequality follows from Lemma 4.3.1. □

We can now define the **branch-informed bounding function**

$$\bar{w}_{\text{bin}}(\mathcal{X}) = \bar{w}(\mathcal{X}) - t_{\bar{0}}(\mathcal{X}, n) \ , \tag{4.3}$$

which has an extra $O(|\mathcal{R}_{\mathcal{X}}|)$ complexity compared to $\bar{w}_{\text{mon}}(\mathcal{X})$. Note that in practice we use both in a chain-like manner, i.e., first evaluate $\bar{w}_{\text{mon}}$ that we get

for free by caching $t_{\bar{\emptyset}}$ after computing $\hat{w}_{\bar{\emptyset}}$, and then proceed with Eq.(4.3) if it fails. We refer to this optimistic estimator as the **chain bounding function** $\bar{w}_{\text{chn}}(\mathcal{X})$.

With these, we use Algorithms 2 and 3 to solve Eq.(4.1). Regarding practicalities, for branch-and-bound we use a priority queue based on potential that leads to the best-first variant. The branching operator $r_{\mathcal{I}}^{H}$ is equivalent to the standard alphabetical enumeration with $r_{\mathcal{I}}^{A}$ (Eq.(3.6)) after initially sorting the input variables in decreasing entropy order. It is important to note that since we sort $\mathcal{I}$ initially, the admissible heuristic of Webb [Web95] to assign the most refinement operators to the least promising nodes (i.e., smallest potential) is not applicable here as it violates the ordering. As $w(\mathcal{X})$ is undefined for $|\mathcal{X}| \leq 1$, we define potential 1 for $|\mathcal{X}| = 1$, and a score of 0 for $|\mathcal{X}| \leq 1$. Moreover, the enumeration order allows for an efficient incremental calculation of $\hat{w}_{\bar{\emptyset}}$.

4.4 Evaluation

In this section we empirically evaluate the proposed discovery framework for dependent sets. In particular, 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 functions, as well as provide concrete findings in example exploratory tasks.

4.4.1 Estimator performance

Here we evaluate the performance of the estimators discussed in this section, i.e., the robust $\hat{w}_0, \hat{w}_{\bar{0}}, \hat{w}_{\bar{\bar{0}}}$ proposed, and the plugin $\hat{w}_{\text{pl}}$. For this evaluation, we first 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 a domain size of 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. Note that due to the varying marginal entropies H of the normalizer, it is not guaranteed that the full (original) joint has the highest w, but rather that the maximum is at least as large.

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 dependency.[5] We sample one distribution for each combination, resulting in 12 different distributions $p^{(i)}, i = 1, \ldots, 12$. We consider data sizes $n \in \{10, 20, 30, \ldots, 100\}$, 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 $n \in \{10, \ldots, 100\}$, since the probability distributions we consider are "small" in size. It is expected, given that all estimators are consistent, that their behavior carries on for larger sample sizes and distributions.

We choose regret to evaluate the estimators as it is an accurate summary of essential properties for an estimator, such as 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 $\tau \in \{\hat{w}_{\text{pl}}, \hat{w}_0, \hat{w}_{\bar{0}}, \hat{w}_{\bar{\bar{0}}}\}$, for which we use exhaustive search to obtain.[6] 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 Figure 4.3 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 much better than the plugin. 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 $\hat{w}_{\bar{0}}$.

Next, in Figure 4.4 we plot the regrets averaged for the different dimensionalities of the joint probability distributions, i.e., $r_n(\tau, \mathcal{P}^2_{[0.1,0.5]})$ (left), $r_n(\tau, \mathcal{P}^3_{[0.1,0.5]})$ (mid-

[5]Note that randomly sampling joint distributions with high normalized total correlation, e.g., in $[0.5, 1]$, is in practice hard for increasing dimensionalities since it requires that all conditional distributions are highly peaked. In addition, this range is less challenging for estimators as it is easily separated from noise.

[6]The $d + 3$ variables are the input variables, the rows are the samples, and an estimator is used as the function to be optimized.

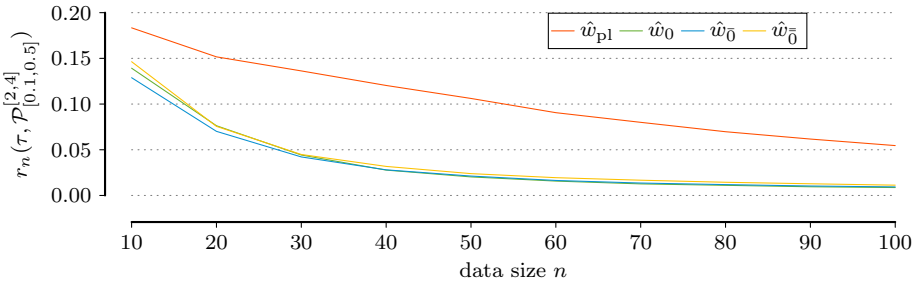


Figure 4.3: Average regret over all models. Regret $r_n(\tau, \mathcal{P}^{[2,4]}_{[0.1,0.5]})$ for sample sizes $n \in \{10, \ldots, 100\}$ and estimators τ.

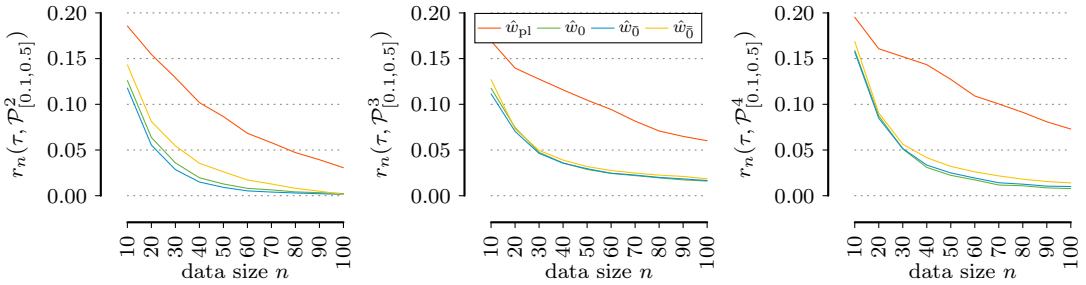


Figure 4.4: Regret curves averaged over different dimensionalities. Average regret $r_n(\tau, \mathcal{P}^2_{[0.1,0.5]})$ (**left**), $r_n(\tau, \mathcal{P}^3_{[0.1,0.5]})$ (**middle**), and $r_n(\tau, \mathcal{P}^4_{[0.1,0.5]})$ (**right**), for sample sizes $n \in \{10, \ldots, 100\}$ and estimators τ.

dle), and $r_n(\tau, \mathcal{P}^4_{[0.1,0.5]})$ (right). Under this different view, we see that the plugin estimator $\hat{w}_{\text{pl}}$ 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 $\hat{w}_{\bar{0}}$ having worse performance. This artifact can be attributed to the following behavior. For small n, not all 5 random variables (2 dependent, 3 independent) get to have samples with domain size 3, and hence, $\hat{w}_{\bar{\bar{0}}}$ that penalizes with the product of domain sizes misses the 2 dependent variables when they are sampled with domain size 3, but the independent ones with domain size 2. In addition, for $d = 2$ the maximum is obtained for the pair of the dependent variables, with its subsets having a score of 0 (since they are singletons). We do not observe this behavior for $d = 3, 4$, for the simple fact that the subsets have a non-zero score,

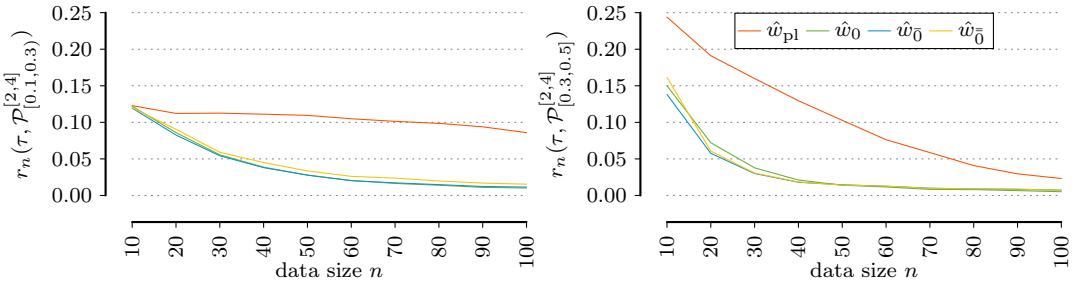


Figure 4.5: Regret curves averaged over "low" and "high" dependency. Average regret $r_n(\tau, \mathcal{P}^{[2,4]}_{[0.1,0.3)})$ (**left**), and $r_n(\tau, \mathcal{P}^{[2,4]}_{[0.3,0.5]})$ (**right**), for sample sizes $n \in \{10, \ldots, 100\}$ and estimators τ.

hence contributing to better regret.

Finally, in Figure 4.5 we plot the regrets averaged over two degrees of dependency, low with $p^{(i)} \in r_n(\tau, \mathcal{P}^{[2,4]}_{[0.1,0.3)})$ (left) and relatively high $p^{(i)} \in r_n(\tau, \mathcal{P}^{[2,4]}_{[0.3,0.5]})$ (right). Again, the corrected estimators have better regret curves. Since their correction is based on a null hypothesis model, they are particularly well-suited for the scenario where the dependency is low, i.e., closer to independence. The plugin $\hat{w}_{\mathrm{pl}}$ on the other hand, cannot distinguish between the chance effects, and hence, has an almost flat curve as we see in the left plot. However, even for better separation with such effects, the corrected estimators still outperform the plugin.

Overall, we see that our proposed robust estimators $\hat{w}_0, \hat{w}_{\bar{0}}$, and $\hat{w}_{\bar{\bar{0}}}$, clearly outperform the plugin, sometimes even by a factor of 5. In addition, we observe that the efficiently computable $\hat{w}_{\bar{\bar{0}}}$ is on par with $\hat{w}_0$ and $\hat{w}_{\bar{0}}$.

4.4.2 Optimization performance

In this section we investigate the performance of the chain bounding function $\bar{w}_{\mathrm{chn}}$ and algorithms proposed for exhaustive and heuristic search for the permutation normalized total correlation $\hat{w}_{\bar{\bar{0}}}$. For the evaluation, we consider benchmark data from the KEEL data repository [SRAFFH+11], 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 4.1. All metric attributes are discretized in 5 equal-frequency bins. This experiment is executed on a Intel Xeon

E5-2643 v3 with 256 GB memory. Our code is online for research purposes.[7]

We employ the two algorithms in order to retrieve the top dependency. For OPUS, 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 4.1 the runtime, the percentage of the pruned search space,[8] the depth of the solution, the maximum depth OPUS had to selectively reach, and the quality $\hat{w}_{\bar{\emptyset}}$ of the top dependent set. For GRD we report runtime and the difference of the quality for the top result with that from OPUS. We average runtimes over 3 independent executions.

We observe that OPUS is highly efficient as it finds the optimum solution in ≤ 30 minutes (i.e., $\alpha = 1$) 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 $\bar{w}_{\text{chn}}$ is very effective in pruning, enabling the discovery of optimum solutions on datasets such as *coil2000* and *movement-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 $\bar{w}_{\text{chn}}$ and enumerated search space allows to selectively explore based on the structure of the data, and not simply by cardinality. That is, it can potentially go to higher levels for promising candidates.

The GRD 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 GRD are almost optimal considering that there are only 2 negligible cases where the two algorithms differ.

Overall, both algorithms are very effective with $\bar{w}_{\text{bin}}$ and $\bar{w}_{\text{mon}}$ as bounding functions. The OPUS algorithm would be preferable in scenarios were solution guarantees are required, while GRD when efficiency is more important, e.g., on very large datasets.

4.4.3 EXAMPLE DISCOVERIES

Last, we proceed with presenting concrete discoveries on three scenarios: finding dependencies on the Tic-tac-toe game, identifying sets of co-inhabitant European

[7] https://github.com/pmandros/fodiscovery

[8] Defined as $100 - (100 * q)/2^d$, where q are the nodes OPUS explored.

land mammals together with factors affecting their coherence, and exploring Bayesian networks.

Tic-tac-toe. First we consider the Tic-tac-toe game (Sec. 3.5.4). Here we treat the 9 attributes $X_1, \ldots, X_9$ corresponding to the cell symbols and the win/loss attribute Y as the input variables $\mathcal{I} = \{X_1, \ldots, X_{10}\}$ (i.e., X_{10} is Y). We present in Figure 4.1 the top-9 results retrieved with $\hat{w}_{\bar{\emptyset}}$. The input variables $X_i, i \in [1, 9]$ 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, $\hat{w}_{\bar{\emptyset}}$ identifies interesting "red" and "blue" dependent sets that can act as latent factors for win and loss, respectively.

Regarding X_{10}, we should be expecting dependency with the losing configurations in a similar manner as the winning ones. This can be attributed to the fact that the losing configurations are in general more random compared to winning, and this combined with the small size of the dataset, cannot support a "losing" top result of size 4.

As a further experiment, we use estimators $\hat{w}_{\text{pl}}, \hat{w}_0, \hat{w}_{\bar{\emptyset}}$ with exhaustive search. We report that $\hat{w}_{\text{pl}}$ 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 $\hat{w}_{\bar{\emptyset}}$ there is agreement with the top 4 of $\hat{w}_{\bar{\emptyset}}$, but the next 5 are all supersets of the top 2 with an extra cell. We find the results of $\hat{w}_{\bar{\emptyset}}$ to be more interesting in this case.

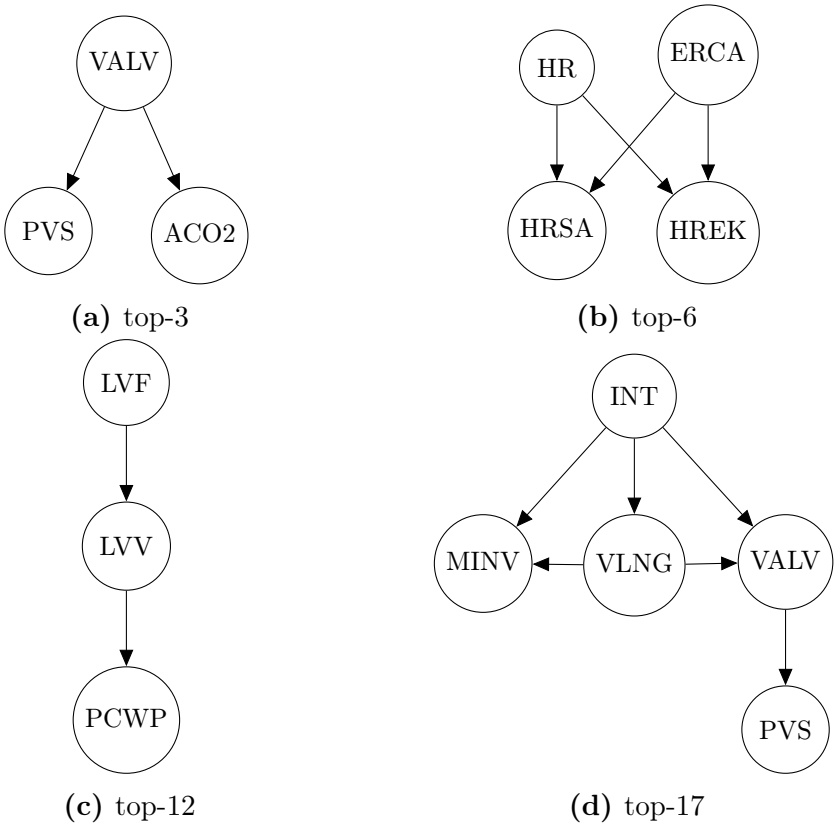
Lastly, we note that the nature of this game implies that the cells are independent, i.e., $p(X_1, \ldots, X_9) = \prod_1^9 p(X_i)$, and that subsets of these cells should become dependent the moment they are conditioned on X_{10}. However, they can take any of 3 values and hence, any dependency is expected to be small. For example, the top-1 of $\hat{w}_{\bar{\emptyset}}$ has score 0.08, and when measured with the plugin $\hat{w}_{\text{pl}}$, has a score of 0.12. These two values are more indicative for the maximum amount of dependency we should expect, in contrast to the value 0.36 for the top-1 retrieved

with $\hat{w}_{\text{pl}}$. To put it differently, $\hat{w}_{\bar{\emptyset}}$ is able to identify aspects of the "low" signal residing in this dataset.

European land mammals. We now shift our attention into data that contain a lot more information, and particular the European land mammal dataset [HFEM07]. The dataset contains presence/absence records of 124 land mammals for a set of 2183 grid cells covering Europe, where each cell is approximately 50×50 km. The dataset also contains enviromental information, such as temperature, precipitation, and elevation, which we discretize into 2 categories to reflect low and high.

In the top results we mainly recover coherent sets of mammals that are categorized as small, i.e., in the families of Insectivora, Rodentia, and Lagomorpha, and are endemic in southern Europe and the European Alps. For example, the top-1 set with score 0.7 contains the Cretan spiny mouse and the Cretan shrew, and top-2 with same score the Savi's pine vole and Crested porcupine, both rodents inhabiting Italy. Larger sets include various species of shrews and rodents. Particularly interesting is the set of the greater white-toothed shrew, the Canarian shrew, and the Osorio shrew. The latter two appear mainly in the Canary islands, while the former in central-west Europe. This set could be used, for example, as an indicator that Osorio shrew, originally described as a separate species, indeed belongs to the shrew family [MBSP03]. Furthermore, we find that the coherence of sets with large mammals depends on the presence of environmental information. As an example, a set with score 0.45 contains two large mammals, moose and Arctic fox, along with three rodents, wood lemming, Norway lemming, and gray red-backed vole. All these inhabit Scandinavia. More coherent sets of large mammals appear together with environmental information, e.g., the set temperature, moose, European bison, and wild goat, with score 0.37. We find that our analysis is to a large extend in sync with that of Heikinheimo et al., and particular the coherent sets of small mammals in southern/central Europe, and the environmental effect on the coherence of sets with large mammals [HFEM07].

Alarm network. Last, we consider the Alarm dataset with $n = 10000$ data samples (Sec. 3.5.3). The goal of this evaluation is to investigate the type of network structures discovered. Ideally, we would like to obtain dependencies that corresponds to connected variable subgraphs. We run the algorithm for the top-30,



(a) top-3 **(b)** top-6

(c) top-12 **(d)** top-17

Figure 4.6: Example discoveries on the Alarm dataset and their corresponding graphical structure.

and report that we discover 5 dependencies of size two, 7 of size three, 10 of size four, and 8 of size five. We present a few in Figure 4.6, and observe that dependencies indeed correspond to connected subgraphs.

4.5 Discussion and conclusions

We considered the problem of measuring and efficiently discovering dependent sets from data. We adopted an information-theoretic approach, and proposed a robust and efficient estimator for normalized total correlation. In addition, we derived two bounding functions to be used for pruning, and proposed effective algorithms for exact, approximate, and heuristic optimization. The results showed

that the estimator has attractive statistical properties, the bounding functions lead to effective optimization algorithms, while qualitative experiments validated that the discoveries are indeed informative.

4.5.1 Different formulations

Note that we can employ total correlation as a global objective: given $\mathcal{I} = \{X_1, \ldots, X_d\}$, discover a decomposition of $\mathcal{I}$ into disjoint sets of variables such that the total correlation is maximized. That is, cluster the attributes into sets of mutually dependent variables. One here could potentially also consider overlapping sets, as a form of soft clustering. This global problem is, however, more computationally demanding for an exact solution. One idea for hard clustering would be to use our current approach to discover the top-1, then remove it and restart for the next top-1. Alternatively, one could employ the conditional total correlation for subsequent iterations and condition on previous solutions.

4.5.2 Future work

Similar to the fraction of information (Ch. 3), we observe again the greedy algorithm performing near-optimal, and similar to Sec. 3.6.1, investigating set function optimization can tell us why. It would be also interesting to consider alternative normalizers for total correlation, e.g., normalize by the cardinality of the set. That way we can favor different types of structures for set dependencies, e.g., larger sets.

Table 4.1: Datasets used in Section 4.4.2. The α values correspond to the maximum possible approximation guarantee in increments of 0.05 such that branch-and-bound (OPUS) finishes in less than 30 minutes. Maximum search level is the maximum level that OPUS had to selectively reach in order to find the solution, while solution depth is the depth where the solution was found. Pruning percentage is the amount of search space reduced by the bounding function and OPUS. The last two columns correspond to the value of the top solution of OPUS, and the difference with the value of the top solution by GRD, respectively.

				search level			time(s)		$\hat{w}_{\bar{\emptyset}}(\mathcal{X}^*)$	
dataset	#rows	#attr.	α	max	sol.	prune%	OPUS	GRD	OPUS	OPUS − GRD
abalone	4174	9	1	6	2	48.90	0.5	0.2	0.67	0
appendic.	106	8	1	3	2	71.37	0.1	0.1	0.56	0
australian	690	15	1	3	2	99.67	0.1	0.1	0.97	0
bupa	345	7	1	5	2	15.70	0.1	0.1	0.10	0
car	1728	7	1	5	2	14.87	0.1	0.1	0.20	0
chess	3196	37	1	9	3	99.99	617.4	0.6	0.64	0
coil2000	9822	86	1	3	2	99.99	7.2	6.7	0.99	0
connect	67557	43	0.8	6	2	99.99	1094.8	11.5	0.62	0
contracept.	1473	10	1	6	2	50.59	0.3	0.1	0.25	0
fars	100968	30	1	2	2	99.99	15.4	10.3	0.99	0
flare	1066	12	1	4	2	93.36	0.1	0.1	0.62	0
german	1000	21	1	6	2	98.63	15.8	0.1	0.26	0
glass	214	10	1	5	2	58.57	0.1	0.1	0.19	0
heart	270	14	1	5	2	83.33	0.4	0.1	0.17	0
ionosphere	351	34	1	5	2	99.99	69.8	0.1	0.45	0
kddcup	494020	42	1	4	2	99.99	284.4	73.5	0.98	0
kr-vs-k	28056	7	1	5	3	8.26	1.6	0.3	0.18	0

led7digit	500	8	1	6	2	37.50	0.1	0.1	0.50	0
letter	20000	17	1	8	2	80.37	390.2	1.2	0.41	0
lymph.	148	19	1	6	2	99.15	0.5	0.1	0.28	0
magic	19029	11	1	5	2	81.63	2.5	0.3	0.67	0
monk	432	7	1	4	2	32.23	0.1	0.1	0.31	0
move. libras	360	91	1	3	2	99.99	12.7	0.5	0.92	0
nursery	12690	9	1	4	2	68.19	0.6	0.2	0.60	0
optdigits	5620	65	0.35	2	2	99.99	3.3	3.4	0.49	0
pageblocks	5472	11	1	5	2	77.71	0.8	0.1	0.69	0
penbased	10992	17	1	7	3	85.38	118	0.8	0.51	0
poker	1025010	11	0.9	8	4	4.95	1760.8	20.6	0.02	0
ring	7400	21	0.1	4	2	99.93	4.4	0.4	0.08	0
saheart	462	10	1	5	2	52.95	0.1	0.1	0.21	0
satimage	6435	37	0.65	6	4	99.99	632.8	1.6	0.55	0.004
segment	2310	20	1	5	2	99.71	2.4	0.1	0.82	0
shuttle	58000	10	1	7	4	57.00	16.2	1.4	0.58	0
sonar	208	61	1	5	2	99.99	1246	0.2	0.35	0
spambase	4597	58	1	4	2	99.99	130.6	2.0	0.89	0
spectf.	267	45	1	5	2	99.99	331.9	0.1	0.29	0
splice	3190	61	0.25	2	2	99.99	1.4	1.5	0.25	0
texture	5500	41	1	3	2	99.99	1.4	1.4	0.99	0
thyroid	7200	22	1	6	2	99.67	26.5	0.5	0.40	0
tic-tac-toe	958	10	1	7	4	11.04	0.4	0.1	0.08	0.005
twonorm	7400	21	0.2	6	2	99.13	84.1	0.4	0.13	0
vehicle	846	19	1	4	2	99.79	0.4	0.1	0.87	0

vowel	990	14	1	2	2	99.43	0.1	0.1	0.95	0
wdbc	569	31	1	4	2	99.99	0.9	0.2	0.90	0
wine	178	14	1	4	2	93.19	0.1	0.1	0.48	0
wine-red	1599	12	1	6	2	53.13	2.1	0.1	0.25	0
wine-white	4898	12	1	7	3	51.29	6.0	0.3	0.32	0
yeast	1484	9	1	5	2	64.21	0.1	0.1	0.19	0
zoo	101	17	1	4	2	99.87	0.1	0.1	0.79	0
avg.	39000	25	0.92	5	2.2	77.00	142	3		

5

Functional Dependency Discovery from Mixed-Type Data

In many practical scenarios data are high-dimensional collections of mixed variable types (i.e., nominal, ordinal, continuous), and estimating the mutual information from such data in a non-parametric way is not trivial. For example, instead of directly considering the underlying continuous variables, we have to resort to their approximations from either data-based discretization or density estimation, with potential loss of information. Moreover, it is not clear how we can do this in the presence of discrete data. The situation becomes even more problematic when we have to efficiently identify the strongest and most robust dependencies for FDD by comparing the estimates for all possible subsets $\mathcal{X} \subseteq \mathcal{I}$.

Proposed mutual information estimators consider mainly the purely discrete and continuous cases. The different families include the discrete, e.g., the plugin $\hat{I}_{\text{pl}}$, the chi-square $\hat{I}_{\chi,\alpha}$, the minimax $\hat{I}_{\text{mm}}$, while for continuous there is adaptive partitioning [SN10, DV99], k-NN [KSG04, BSY19], and kernel density estimation [PY08, GSG15]. For mixed data, the state-of-the-art k-NN [GKOV17] based

This chapter is an extended version of work that originally appeared in ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) [MKBV20].

on the Radon-Nikodym derivative is applicable for multivariate mixtures. None of the above, however, fits to our mixed data FDD scenario. The continuous estimators are defined for Euclidean spaces, where nominal attributes cannot be trivially embedded. Moreover, given purely discrete data, the Radon-Nikodym mixed estimator recovers the plugin estimator $\hat{I}_{\text{pl}}$ that trivially fails the FDD task. Discrete estimators, on the other hand, can work with continuous data after discretization has been applied. While efficiently discovering robust dependencies in discrete data has been principally addressed in the previous chapters, it remains unclear with what quantization methods it can be combined such that the search consistently identifies the strongest dependencies in mixed-type data. We solve these with the following contributions:

- first, to arrive at a consistent mixed estimator $\hat{I}_{mx}$, we recall that mutual information for two continuous random variables can be attained as a limit along a refining quantization sequence [CT06, Sec. 8.3]. We extend this result for mixed sets of variables, as well as identify the class of quantizations applicable that includes known techniques such as equal-frequency. We then translate this process to empirical samples, and identify the requirements for consistency (Sec. 5.2).
- Second, based on the theory developed we propose a framework for mixed mutual information estimation and demonstrate how it can be applied in practice for FDD (Sec. 5.3).
- Third, we combine the mixed estimator with the robust FDD framework developed in Chapter 3. In particular, we show that the permutation mutual information estimator is well-suited for the mixed estimator framework, and modify the algorithms for exact, approximate, and heuristic search (Sec. 5.4, see Fig. 5.1 for a demonstration).
- Lastly, we perform extensive evaluation on a wide range of real and synthetic data (Sec. 5.5).

We start with preliminaries in Section 5.1, and end with discussion and conclusions in Section 5.6.

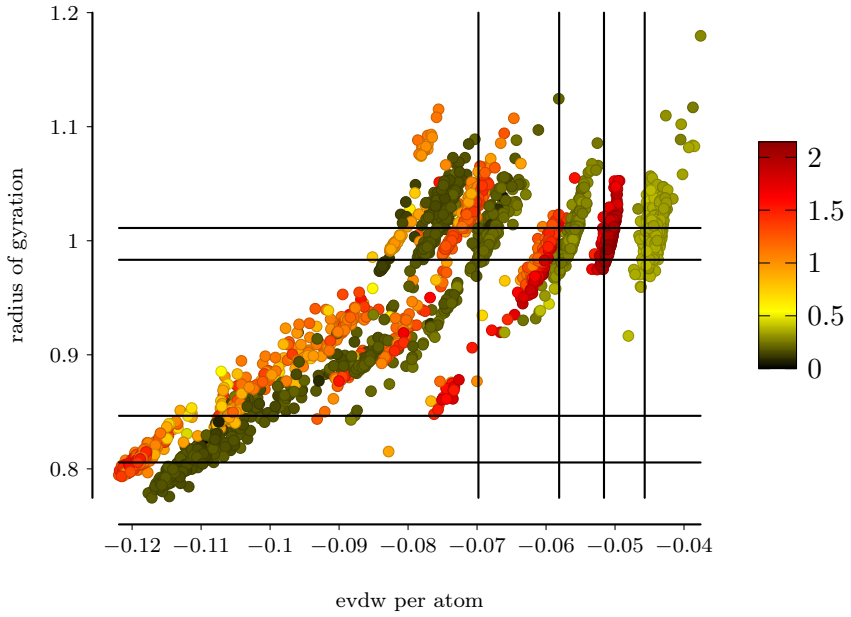


Figure 5.1: Functional dependency discovery from mixed data on a Materials Science case study. The dataset contains 12200 gold cluster configurations (of sizes 5 to 14 atoms) generated at finite temperature by replica-exchange molecular dynamics simulations [GBV+17]. The attributes in this dataset are 23 physicochemical and geometrical properties of the gold clusters. Here we are interested in discovering dependencies that are descriptive for the target variable HOMO-LUMO gap that determines the electro-chemical properties of a cluster. Out of all possible $2^{22} - 1$ variable subsets, our proposed mixed-data FDD method uncovers that structural variable "radius of gyration" and non-local dispersion energies "evdw per atom" approximately determine the target with $\hat{F}_0$ score 0.43. The scatterplot represents the nano-clusters against the two-dimensional descriptor, with color indicating the values of the HOMO-LUMO gap. Black lines represent the resulting partition in $\mathbb{R}^2$ with a budget of up to 5 bins per axis (cOP, $l = 5, c = 2$).

5.1 Preliminaries

We often use $\mathcal{D}, \mathcal{G}$, to indicate sets of discrete variables, and $\mathcal{C}$ for sets of continuous variables. We consider **quantization strategies** for continuous random variables which we denote with Q. Given $k \in \mathbb{Z}^+$ and a continuous random variable C, Q produces a partition $Q_k = \{S_1, \ldots, S_k\}$ of the domain $V_C \subseteq \mathbb{R}$ in k consecutive intervals with $\cup_{i=1}^{k} S_i = V_C$ (upper-bound exclusive). With C_{Q_k} we represent the

quantized C according to Q and k. As an example, **equal-frequency** denoted as Q^{EF}, partitions C with $Q_k^{\text{EF}} = \{S_1, \ldots, S_k\}$ such that $\int_{S_i} f_C(c)dc = 1/k$ for all $i \in [k]$, where $f_C(c)$ is the density function of C. Given Q and k, we use δ_i for the corresponding length of the sub-interval S_i. In this paper, we are interested in the class of quantization strategies for which $\max_{i\in[k]} \delta_i \to 0$ as $k \to \infty$, which we refer to as **converging strategies**. These notions extend to the multivariate case $\mathcal{C} = \{C_1, \ldots, C_m\}$, with $Q_{k^m} = \{\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{k^m}\}$ being a partition of $V_{\mathcal{C}} \subseteq \mathbb{R}^m$, produced by partitioning each $C \in \mathcal{C}$ in k bins. We use Q_k whenever clear from the context. For a Q, the set $\Pi_l(Q) = \{Q_1, \ldots, Q_l\}$ corresponds to all partitions by Q in up to l bins, and $\Pi_l^m(Q)$ to the set of all partitions for domains in $\mathbb{R}^m$.

We define the following relation for two partitions: Q'_v is a **refinement** of Q_u, denoted as $Q_u \preceq Q'_v$, if $v \geq u$ and there exists a map $r : [u] \to 2^{[v]}$, such that for every $i \in [u]$, we have $S_i = \cup_{j\in r(i)} S'_j$. For example, we have that $Q_2^{\text{EF}} = \{S_1, S_2\} \preceq Q_4^{\text{EF}} = \{S'_1, S'_2, S'_3, S'_4\}$, since $S_1 = S'_1 \cup S'_2$ and $S_2 = S'_3 \cup S'_4$.

Given samples, a quantization strategy Q translates to a **discretization strategy**, denoted as $\hat{Q}$, that corresponds to the same strategy to partition the n sample points X_s in k bins, where X_s is X **sorted in ascending order**. For example, let us consider random variable $X \sim \mathbf{U}(-1, 1)$, and a sorted sample $X = [-0.5, -0.3, 0, 0.6, 0.9, 1]$. For $k = 3$ and equal-frequency, $\hat{\pi} = \hat{Q}_3^{\text{EF}}$ can be seen as a map $\hat{\pi} : \mathbb{R} \to \{1, 2, 3\}$ that splits the data sample in three bins of two points each, to create discrete variable $X_{\hat{\pi}} = [1, 1, 2, 2, 3, 3]$ with domain $V_{X_{\hat{\pi}}} = \{1, 2, 3\}$. With $\Pi_{l,n}$, we denote the set of all possible partitions of n data points in up to $l \leq n$ bins, and for a $\hat{Q}$, we have $\Pi_{l,n}(\hat{Q}) = \{\hat{Q}_1, \ldots, \hat{Q}_l\}$. Note that we also consider X_π for $\pi = Q_k^{\text{EF}}$, meaning that X is discetized according to the equal-frequency quantization of the population domain $V_X = [-1, 1]$, that is, for $\pi = Q_3^{\text{EF}} = \{[-1, -1/3), [-1/3, 1/3), [1/3, 1]\}$, $X_\pi = [1, 2, 2, 3, 3, 3]$.

Finally, recall the notion of **dominated convergence**: let a_{mn} be a sequence such that for all m the limit $a_m^* = \lim_{n\to\infty} a_{mn}$ exists. Further, let $p_m \geq 0$ be another sequence and let $u_m \geq |a_{mn}|$ for all m, n such that $\sum_m p_m u_m < \infty$. Then the limit $\lim_{n\to\infty} \sum_m p_m a_{mn}$ exists and is equal to $\sum_m p_m a_m^*$.

5.2 Consistent mixed mutual information estimation

In this section we introduce the information-theoretic notions of multivariate entropy and mutual information for mixtures of discrete (both nominal and ordinal) and continuous random variables. We demonstrate how a sequence of finer-grained quantizations of continuous random variables leads to the actual (i.e., unquantized) mutual information. Finally, we show how this process translates to empirical samples, enabling estimation from mixed-type data.

Given sets $\mathcal{D}$ and $\mathcal{C}$ of discrete and continuous random variables, respectively, the **Shannon entropy** of $\mathcal{D} \cup \mathcal{C}$ with joint probability distribution $f(\mathbf{d}, \mathbf{c}) = f_{\mathcal{C} \mid \mathbf{d}}(\mathbf{c} \mid \mathbf{d}) p(\mathbf{d})$ is

$$\begin{aligned} H(\mathcal{D}, \mathcal{C}) = & - \sum_{\mathbf{d} \in \mathcal{D}} \int_{\mathcal{C}} f(\mathbf{d}, \mathbf{c}) \log f(\mathbf{d}, \mathbf{c}) d\mathbf{c} \\ = & - \sum_{\mathbf{d} \in \mathcal{D}} p(\mathbf{d}) \int_{\mathcal{C}} f_{\mathcal{C} \mid \mathbf{d}}(\mathbf{c} \mid \mathbf{d}) \log f_{\mathcal{C} \mid \mathbf{d}}(\mathbf{c} \mid \mathbf{d}) d\mathbf{c} \\ & - \sum_{\mathbf{d} \in \mathcal{D}} p(\mathbf{d}) \log p(\mathbf{d}) \int_{\mathcal{C}} f_{\mathcal{C} \mid \mathbf{d}}(\mathbf{c} \mid \mathbf{d}) d\mathbf{c} \\ = & H(\mathcal{C} \mid \mathcal{D}) + H(\mathcal{D}) \ . \end{aligned}$$

Let us consider a converging Q and $Q_k = \{\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{k^m}\}$ an m-dimensional partition of the domain $V_{\mathcal{C}} \subseteq \mathbb{R}^m$. Let us assume that $f_{\mathcal{C} \mid \mathbf{d}}(\mathbf{c} \mid \mathbf{d})$ is continuous within each hypercube for all $\mathbf{d} \in \mathcal{D}$. Then, using the **mean value theorem** for integrals, there exists a value $\mathbf{c}_i$ within each hypercube i such that $f_{\mathcal{C} \mid \mathbf{d}}(\mathbf{c}_i \mid \mathbf{d}) \boldsymbol{\delta}_i = \int_{\boldsymbol{S}_i} f_{\mathcal{C} \mid \mathbf{d}}(\mathbf{c} \mid \mathbf{d}) d\mathbf{c}$. The quantized $\mathcal{C}$ is defined as $\mathcal{C}_{Q_k} = \mathbf{c}_i$ for $\mathcal{C} \in \boldsymbol{S}_i$, and has conditional probability $p_{i \mid \mathbf{d}} = \boldsymbol{\delta}_i f_{\mathcal{C} \mid \mathbf{d}}(\mathbf{c}_i \mid \mathbf{d})$ that $\mathcal{C}_{Q_k} = \mathbf{c}_i$ when $\mathcal{D} = \mathbf{d}$. The following lemma shows how $H(\mathcal{D}, \mathcal{C}_{Q_k})$ converges to $H(\mathcal{D}, \mathcal{C})$.

Lemma 5.2.1. *Given random variables $\mathcal{D}$ of finite domain $V_{\mathcal{D}}$, random variables $\mathcal{C}$, and converging Q, if the conditional density $f_{\mathcal{C} \mid \boldsymbol{d}}(\boldsymbol{c} \mid \boldsymbol{d})$ is Riemann integrable for all $\boldsymbol{d} \in \mathcal{D}$, then*

$$\lim_{k \to \infty} H(\mathcal{D}, \mathcal{C}_{Q_k}) + \beta_{Q_k}(\mathcal{D}) = H(\mathcal{D}, \mathcal{C}) \ ,$$

where $\beta_{Q_k}(\mathcal{D}) = \sum_{\boldsymbol{d}\in\mathcal{D}} p(\boldsymbol{d}) \sum_i \boldsymbol{\delta}_i f_{\mathcal{C}}(\boldsymbol{c}_i \mid \boldsymbol{d}) \log \boldsymbol{\delta}_i$. *Further, if for all* $\boldsymbol{d} \in \mathcal{D}$, k, *we have* $h_k(\boldsymbol{d}) = \left|\sum_{i=1}^{k} \boldsymbol{\delta}_i f(\boldsymbol{c}_i \mid \boldsymbol{d}) \log f(\boldsymbol{c}_i \mid \boldsymbol{d})\right| \leq a(\boldsymbol{d})$ *such that* $\sum_{\boldsymbol{d}} p(\boldsymbol{d}) a(\boldsymbol{d}) < \infty$, *the result also holds for infinite* $V_{\mathcal{D}}$.

Proof. We write $f_{\mathcal{C}}(\mathbf{c}_i \,|\, \mathbf{d})$ instead of $f_{\mathcal{C}\,|\,\mathbf{d}}(\mathbf{c}_i \,|\, \mathbf{d})$. We have

$$\begin{aligned} H(\mathcal{D}, \mathcal{C}_{Q_k}) &= -\sum_{\mathbf{d}\in\mathcal{D}} p(\mathbf{d}) \sum_{i=1}^{k} \boldsymbol{\delta}_i f_{\mathcal{C}}(\mathbf{c}_i \,|\, \mathbf{d}) \log\left(\boldsymbol{\delta}_i f_{\mathcal{C}}(\mathbf{c}_i \,|\, \mathbf{d})\right) + H(\mathcal{D}) \\ &= -\sum_{\mathbf{d}\in\mathcal{D}} p(\mathbf{d}) \sum_{i=1}^{k} \boldsymbol{\delta}_i f_{\mathcal{C}}(\mathbf{c}_i \,|\, \mathbf{d}) \log f_{\mathcal{C}}(\mathbf{c}_i \,|\, \mathbf{d}) \\ &\quad \underbrace{-\sum_{\mathbf{d}\in\mathcal{D}} p(\mathbf{d}) \sum_{i=1}^{k} \boldsymbol{\delta}_i f_{\mathcal{C}}(\mathbf{c}_i \,|\, \mathbf{d}) \log \boldsymbol{\delta}_i}_{\beta_{Q_k}(\mathcal{D})} + H(\mathcal{D}) \quad . \end{aligned}$$

Since $f_{\mathcal{C}\,|\,\mathbf{d}}(\mathbf{c} \,|\, \mathbf{d})$ is Riemann integrable, the inner sum of the first term converges to its integral as $k \to \infty$. For finite $V(\mathcal{D})$, the first sum then converges to $H(\mathcal{C} \mid \mathcal{D})$. For infinite $V(\mathcal{D})$, the sum also converges to $H(\mathcal{C} \mid \mathcal{D})$ as $\sum_{\mathbf{d}} p(\mathbf{d}) a(\mathbf{d}) < \infty$ is the assumption required for dominated convergence. □

Lemma 5.2.1 states that for convergence a sequence of finer-grained quantizations and a correction by β are required. In addition, $h_k(\mathbf{d})$ have to be bounded for convergence with infinite $V(\mathcal{D})$. Note that the correction $\beta_{Q_k}(\mathcal{D})$ is necessary due to the **infinite quantization error** as $k \to \infty$. That is, as the partitions get finer, $H(\mathcal{D}, \mathcal{C}_{Q_k})$ diverges. We also note that the entropy $H(\mathcal{D}, \mathcal{C})$, unlike the discrete case $H(\mathcal{D})$, can be negative, e.g., for $C \sim \mathbf{U}(0, a), a < 1$ [CT06, Sec. 8.1]. These, however, do not extend to mutual information.

The **mutual information** for $\mathcal{X} = \{\mathcal{D}, \mathcal{C}\}$ and $\mathcal{Y} = \{\mathcal{D}', \mathcal{C}'\}$, is defined as $I(\mathcal{X}; \mathcal{Y}) = H(\mathcal{D}, \mathcal{C}) + H(\mathcal{D}', \mathcal{C}') - H(\mathcal{D}, \mathcal{C}, \mathcal{D}', \mathcal{C}')$, and it holds that $I(\mathcal{X}; \mathcal{Y}) \geq 0$. We proceed with the following theorem about the convergence of $I(\mathcal{X}; \mathcal{Y})$ w.r.t. the quantization process.

Theorem 5.2.1. *Given random variables* $\mathcal{X} = \{\mathcal{D}, \mathcal{C}\}, \mathcal{Y} = \{\mathcal{D}', \mathcal{C}'\}$, *with Riemann integrable conditional density* $f_{\mathcal{C},\mathcal{C}'\,|\,\boldsymbol{d},\boldsymbol{d}'}(\boldsymbol{c}, \boldsymbol{c}' \mid \boldsymbol{d}, \boldsymbol{d}')$ *for all* $\boldsymbol{d} \in \mathcal{D}, \boldsymbol{d}' \in \mathcal{D}'$, *as*

well as converging Q, Q', then

$$I(\mathcal{X};\mathcal{Y}) = \lim_{k\to\infty} I(\mathcal{D},\mathcal{C}_{Q_k};\mathcal{D}',\mathcal{C}'_{Q'_k}) \ .$$

Proof. For readability, we drop k, as well as use $f_{\mathcal{C}}(\mathbf{c}_i \,|\, \mathbf{d})$ instead of $f_{\mathcal{C}\,|\,\mathbf{d}}(\mathbf{c}_i \,|\, \mathbf{d})$, whenever clear from the context. We have:

$$\begin{aligned}
&I(\mathcal{D},\mathcal{C}_Q;\mathcal{D}',\mathcal{C}'_{Q'}) = H(\mathcal{D},\mathcal{C}_Q) + H(\mathcal{D}',\mathcal{C}'_{Q'}) - H(\mathcal{D},\mathcal{C}_Q,\mathcal{D}',\mathcal{C}'_{Q'}) \\
&= -\sum_{\mathbf{d}\in\mathcal{D}} p(\mathbf{d}) \sum_i \boldsymbol{\delta}_i f_{\mathcal{C}}(\mathbf{c}_i \,|\, \mathbf{d}) \log f_{\mathcal{C}}(\mathbf{c}_i \,|\, \mathbf{d}) + \beta_Q(\mathcal{D}) + H(\mathcal{D}) \\
&\quad - \sum_{\mathbf{d}'\in\mathcal{D}'} p(\mathbf{d}') \sum_j \boldsymbol{\delta}'_j f_{\mathcal{C}'}(\mathbf{c}_j \,|\, \mathbf{d}') \log f_{\mathcal{C}'}(\mathbf{c}_j \,|\, \mathbf{d}') + \beta_{Q'}(\mathcal{D}') + H(\mathcal{D}') \\
&\quad + \sum_{\mathbf{d}\in\mathcal{D},\mathbf{d}'\in\mathcal{D}'} p(\mathbf{d},\mathbf{d}') \sum_{i,j} \boldsymbol{\delta}_i \boldsymbol{\delta}'_j f_{\mathcal{C},\mathcal{C}'}(\mathbf{c}_i,\mathbf{c}_j \,|\, \mathbf{d},\mathbf{d}') \log f_{\mathcal{C},\mathcal{C}'}(\mathbf{c}_i,\mathbf{c}_j \,|\, \mathbf{d},\mathbf{d}') \\
&\quad - \beta_{Q,Q'}(\mathcal{D},\mathcal{D}') - H(\mathcal{D},\mathcal{D}') \ .
\end{aligned}$$

We know from Lemma 5.2.1 that the sums converge to $H(\mathcal{C} \mid \mathcal{D}), H(\mathcal{C}' \mid \mathcal{D}')$, and $H(\mathcal{C},\mathcal{C}' \mid \mathcal{D},\mathcal{D}')$. It remains to show that $\beta_{Q,Q'}(\mathcal{D},\mathcal{D}') = \beta_Q(\mathcal{D}) + \beta_{Q'}(\mathcal{D}')$. We have

$$\begin{aligned}
\beta_{Q,Q'}(\mathcal{D},\mathcal{D}') &= \sum_{\mathbf{d}\in\mathcal{D}} \sum_{\mathbf{d}'\in\mathcal{D}'} p(\mathbf{d},\mathbf{d}') \sum_{i=1} \sum_{j=1} \boldsymbol{\delta}_i \boldsymbol{\delta}'_j f_{\mathcal{C},\mathcal{C}'}(\mathbf{c}_i,\mathbf{c}_j \,|\, \mathbf{d},\mathbf{d}') \log(\boldsymbol{\delta}_i \boldsymbol{\delta}'_j) \\
&= \sum_{\mathbf{d}\in\mathcal{D}} \sum_{\mathbf{d}'\in\mathcal{D}'} p(\mathbf{d},\mathbf{d}') \sum_{i=1} \boldsymbol{\delta}_i f_{\mathcal{C}}(\mathbf{c}_i \,|\, \mathbf{d},\mathbf{d}') \log(\boldsymbol{\delta}_i) \\
&\quad + \sum_{\mathbf{d}\in\mathcal{D}} \sum_{\mathbf{d}'\in\mathcal{D}'} p(\mathbf{d},\mathbf{d}') \sum_{j=1} \boldsymbol{\delta}'_j f_{\mathcal{C}'}(\mathbf{c}_j \,|\, \mathbf{d},\mathbf{d}') \log(\boldsymbol{\delta}'_j) \ .
\end{aligned}$$

Let us focus on the first term, for which we have

$$\begin{aligned}
\sum_{\mathbf{d}\in\mathcal{D}} p(\mathbf{d}) \sum_{i=1} \boldsymbol{\delta}_i \log(\boldsymbol{\delta}_i) \sum_{\mathbf{d}'\in\mathcal{D}'} p(\mathbf{d}' \,|\, \mathbf{d}) f_{\mathcal{C}}(\mathbf{c}_i \,|\, \mathbf{d},\mathbf{d}') &= \sum_{\mathbf{d}\in\mathcal{D}} p(\mathbf{d}) \sum_{i=1} \boldsymbol{\delta}_i \log(\boldsymbol{\delta}_i) f_{\mathcal{C}}(\mathbf{c}_i \,|\, \mathbf{d}) \\
&= \beta_Q(\mathcal{D}) \ .
\end{aligned}$$

Similarly, the second term is $\beta_{Q'}(\mathcal{D}')$, and therefore $\beta_{Q,Q'}(\mathcal{D},\mathcal{D}') = \beta_Q(\mathcal{D}) + \beta_{Q'}(\mathcal{D}')$,

concluding the proof. □

Theorem 5.2.1 states that the unquantized $I(\mathcal{X};\mathcal{Y})$ is attained for converging strategies Q. We now proceed to translate this quantization process for samples of $p(\mathcal{X},\mathcal{Y})$, enabling the estimation from mixed data in practice. For this, we use consistent discrete estimators $\hat{H}$ for entropy H and their corresponding sampling complexities $S_{\hat{H}}$.

Recall that an estimator $\hat{H}$ is called consistent (Sec. 2.2) if $\hat{H} \xrightarrow{p} H$ as $n \to \infty$. For entropy, the sample complexity, i.e., the minimum sample size that achieves a certain concentration (ϵ-δ-PAC guarantee), is usually expressed as a function of the domain size. For example, the plugin $\hat{H}_{\mathrm{pl}}$ has sample complexity $S_{\hat{H}_{\mathrm{pl}}}(k) \in O(k)$ where k the domain size. The main idea of the following theorem is to use consistent estimators for Shannon entropy and upper-bound the number of partitions for a given number of samples n w.r.t. their sample complexity.

Theorem 5.2.2. *Let $\mathcal{X} = \{\mathcal{D}, \mathcal{C}\}$, $\mathcal{Y} = \{\mathcal{D}', \mathcal{C}'\}$ be i.i.d. samples from $p(\mathcal{X},\mathcal{Y})$, with finite $V_{\mathcal{D}}, V_{\mathcal{D}'}$ and Riemann integrable conditional densities $f(\boldsymbol{c}, \boldsymbol{c}' \mid \boldsymbol{d}, \boldsymbol{d}')$. Further, let Q, Q' be two converging strategies, $\hat{H}$ a consistent estimator for discrete entropy, and $g(n)$ a strictly increasing function such that $g(n) \leq S_{\hat{H}}^{-1}(n)$. Then*

$$\lim_{n\to\infty} \hat{I}(\mathcal{D}, \mathcal{C}_{Q_{g(n)}}; \mathcal{D}', \mathcal{C}'_{Q'_{g(n)}}) = I(\mathcal{X},\mathcal{Y}) \ .$$

Further, if $\hat{p}(\boldsymbol{d}, \boldsymbol{d}') \xrightarrow{L1} p(\boldsymbol{d}, \boldsymbol{d}')$ and $\hat{H}(\mathcal{C}_{Q_k} \mid \boldsymbol{d}, \boldsymbol{d}') + \hat{H}(\mathcal{C}'_{Q'_k} \mid \boldsymbol{d}, \boldsymbol{d}') \leq \alpha$ uniformly for all $\boldsymbol{d} \in V_{\mathcal{D}}, \boldsymbol{d}' \in V_{\mathcal{D}'}$ and $k \in \mathbb{Z}^+$, the result also holds for countably infinite $V(\mathcal{D}), V(\mathcal{D}')$.

Proof. We drop subscripts from Q, Q' for readability. Note that the latter two assumptions are implied for finite $V(\mathcal{D}), V(\mathcal{D}')$, and hence, we prove the more general statement. We have

$$\hat{I}(\mathcal{X};\mathcal{Y}) = \hat{H}(\mathcal{D}, \mathcal{C}_Q) + \hat{H}(\mathcal{D}', \mathcal{C}'_{Q'}) - \hat{H}(\mathcal{D}, \mathcal{C}_Q, \mathcal{D}', \mathcal{C}'_{Q'}) \ .$$

Now, let us focus on the first term, i.e., $\hat{H}(\mathcal{D}, \mathcal{C}_Q) = \hat{H}(\mathcal{C}_Q \,|\, \mathcal{D}) + \hat{H}(\mathcal{D})$. For $\hat{H}(\mathcal{D})$,

we know it converges due to the consistency of $\hat{H}$. For $\hat{H}(\mathcal{C}_Q \,|\, \mathcal{D})$, we have

$$\begin{aligned}\hat{H}(\mathcal{C}_Q \mid \mathcal{D}) &= \sum_{\mathbf{d}\in\mathcal{D}} \hat{p}(\mathbf{d})\hat{H}(\mathcal{C}_Q \mid \mathcal{D}=\mathbf{d}) \\ &= \sum_{\mathbf{d}\in\mathcal{D}} \big(\hat{p}(\mathbf{d}) - p(\mathbf{d})\big)\hat{H}(\mathcal{C}_Q \mid \mathcal{D}=\mathbf{d}) \\ &\quad + \sum_{\mathbf{d}\in\mathcal{D}} p(\mathbf{d})\hat{H}(\mathcal{C}_Q \mid \mathcal{D}=\mathbf{d}) \ .\end{aligned}$$

Since all $\hat{H}(\mathcal{C}_{\hat{Q}} \mid \mathcal{D}=\mathbf{d}) \leq \alpha$ are bounded and $\hat{p} \xrightarrow{L1} p$, the first sum converges to zero. For the second sum, we have that $\lim_{n\to\infty} \hat{H}(\mathcal{C}_{\hat{Q}} \mid \mathcal{D}=\mathbf{d}) = H(\mathcal{C} \mid \mathcal{D}=\mathbf{d})$ due to the additional assumption for $\hat{H}$. Hence, the complete sum converges to $H(\mathcal{C} \mid \mathcal{D})$ as the conditions for dominated convergence apply. Analogous arguments for the remaining entropy terms establish the result. □

Theorem 5.2.2 states the two requirements for convergence to $I(\mathcal{X};\mathcal{Y})$ given i.i.d. samples: converging quantization strategies and consistent discrete estimators for entropy. To end this section, we make the following two remarks. In exploratory scenarios with no access to p, a Q that partitions the variable domain is not directly applicable. Instead, we use the empirical $\hat{Q}$. Note, however, that for EF we have $\hat{Q}_k \overset{n\to\infty}{\longrightarrow} Q_k$. In the remainder of this chapter we remove hat symbols for $\hat{Q}$. For the second remark, while the necessary and sufficient requirements for consistent entropy estimation is $S_{\hat{H}}(k) \in \Omega(k/\log(k))$, in this chapter we study "slower" estimators of the form $\hat{I}_{\text{pl}} + b(n)$, with $b(n) \overset{n\to\infty}{\longrightarrow} 0$. The reason is that these estimators are more flexible w.r.t. the FDD task, e.g., $b(n)$ directly penalizes data sparsity and optimization algorithms have been provided. In the next section, we derive a mutual information estimator for mixed variables.

5.3 Practical mixed data estimator

We start by noting that attaining the population value $I(\mathcal{X};\mathcal{Y})$ via quantization can be equivalently formulated as a supremum over all finite partitions of the

domains $V_{\mathcal{C}}, V_{\mathcal{C}'}$. Translating this to a sample, we arrive at an estimator of the form

$$\max_{\boldsymbol{\pi}\in\Pi_{l,n}^{|\mathcal{C}|},\boldsymbol{\pi}'\in\Pi_{l,n}^{|\mathcal{C}'|}} \hat{I}(\mathcal{D},\mathcal{C}_{\boldsymbol{\pi}};\mathcal{D}',\mathcal{C}'_{\boldsymbol{\pi}'}) \ ,$$

i.e., the optimization problem of maximizing a discrete consistent estimator $\hat{I}$ over the set of all possible partitions $\Pi_{l,n}^{|\mathcal{C}|}$ and $\Pi_{l,n}^{|\mathcal{C}'|}$, with $l \in \mathbb{Z}^+$ being the maximum number of bins. For our FDD purposes, we consider the mutual information $I(\mathcal{X};Y)$ between $\mathcal{X} = \{\mathcal{D},\mathcal{C}\}$ and a univariate discrete target Y, i.e., the case

$$\hat{I}_{mx}(\mathcal{X},Y) = \max_{\boldsymbol{\pi}\in\Pi_{l,n}^{|\mathcal{C}|}} \hat{I}(\mathcal{D},\mathcal{C}_{\boldsymbol{\pi}};Y) \ . \tag{5.1}$$

This optimization problem of Eq.(5.1), however, is infeasible in practice: the search space is prohibitively large with $|\mathcal{C}|\sum_{i=0}^{l}\binom{n-1}{i}$ possible $|\mathcal{C}|$-dimensional partitions $\boldsymbol{\pi}$ in up to l bins. Moreover, while estimators $\hat{I}$ are consistent, they can be statistically inefficient for limited data samples and almost trivially produce arbitrary partitions and estimates due to data sparsity in the $|\mathcal{X}|$-dimensional space. We present solutions for both problems, starting with the former.

5.3.1 Optimization

First, let us assume $|\mathcal{C}| = m$, and reformulate the problem. Instead of directly searching for high-dimensional partitions $\boldsymbol{\pi} \in \Pi_{l,n}^{m}$, we can equivalently search for m univariate partitions, i.e.,

$$\max_{\pi_1,\ldots,\pi_m\in\Pi_{l,n}} \hat{I}(\mathcal{D},\{C_{1\pi_1},\ldots,C_{m\pi_m}\};Y) \ .$$

This approach allows us to consider the abundant research on partitioning the real line $\mathbb{R}$. Here, we provide two solutions from prior work on dependency estimation. The first has been used for an exact solution, while the second for an approximate. Let us focus for now on the univariate continuous case, i.e., $\mathcal{X} = \{C\}$.

For an exact solution, note that a naive algorithm would perform exhaustive search through all $\sum_{i=0}^{l}\binom{n-1}{i}$ partitions for C. However, Reshef et al. in seminal work on dependency estimation for pairs of continuous variables [RRF+11], give a

polynomial time algorithm for the plugin $\hat{I}_{\text{pl}}$, exploiting the **optimal substructure** of $\max_{\pi \in \Pi_{l,n}} \hat{I}_{\text{pl}}(C_\pi; Y)$: the best partition in up to l bins is comprised of the best partition in up to $l-1$ bins. The **dynamic programming (DP)** algorithm has complexity $O(ln^2)$. For efficiency, the authors propose a relaxation where C is partitioned in l equal-frequency bins, and DP finds the best partition from $\{\pi : \pi \preceq Q_l^{\text{EF}}\}$. For more candidate partitions, a parameter $c \in \mathbb{Z}^+$ controls the number of initial bins via cl (see [RRF+11, Sup. material, Sec. 3.2.2]). The complexity now is $O(c^2 l^3)$, and we refer to this partitioning scheme as **contrained optimal partition (cOP)**, with $\Pi_{l,n}(Q^{\text{cOP}}) = \{\pi : \pi \preceq Q_{cl}^{\text{EF}}, |\pi| \leq l\}$ for parameter c. For $cl = n$, cOP becomes optimal. The approximate technique is based on **equal-frequency**. To find an appropriate partition for estimating mutual information from pairs of discrete/continuous random variables, Suzuki suggests to pick the equal-frequency partition that maximizes mutual information in up to $l = 0.5 \log_2(n)$ bins, i.e., $\max_{k \in [l]} \hat{I}(C_{Q_k^{\text{EF}}}; Y)$ [Suz19]. Sugiyama and Borgwardt perform the same process in order to estimate the information dimension of a continuous variable, with $l = \log_2(n)$ [SB13]. For Q^{EF}, we have $\Pi_{l,n}(Q^{\text{EF}}) = \{\pi : \pi = Q_k^{\text{EF}}, k \in [l]\}$. Regarding the two techniques, cOP has a clear advantage: a larger space of candidate partitions controlled by parameters based on the availability of resources. However, EF has the negligible complexity of $O(l)$. In addition, EF is applicable to any estimator $\hat{I}$, while cOP requires optimal substructure for the polynomial DP algorithm.

Now given set $\mathcal{X} = \{\mathcal{D}, \mathcal{C}\}$, in order to perform a multidimensional discretization in practice, we adopt a greedy approach of **iteratively discretizing** one $C \in \mathcal{C}$ at a time. Note that while this approach is greedy in nature, the choice for a partition is done jointly with all the already discrete and discretized variables. In addition, the consistency should not violated for $k, n \to \infty$. Since the result now depends on the order, we first sort the variables $X \in \mathcal{X}$ in decreasing order of **marginal mutual information** $\hat{I}(X; Y)$. For the continuous $C \in \mathcal{C}$, we marginally discretize them according to Q and l. That way, we let the most informative continuous variables discretize first, jointly with the already discrete. The details of our proposed mixed estimator framework are shown in Algorithm 5. Given mixed set of random variables $\mathcal{X} = \{\mathcal{D}, \mathcal{C}\}$, discrete target Y, partitioning

Algorithm 5 $\hat{I}_{mx}$: Given set of mixed random variables $\mathcal{X} = \{\mathcal{D}, \mathcal{C}\}$, discrete target Y, partitioning strategy Q, consistent discrete estimator $\hat{I}$, and maximum number of bins l, the algorithm returns an estimate of $I(\mathcal{X}; Y)$

function $\hat{I}_{mx}(\mathcal{X}, Y, Q, \hat{I}, l)$
  $\mathcal{X}' = \text{sortMarginally}(\mathcal{X}, Y, Q, \hat{I}, l)$
  $\mathcal{G} = \emptyset$
  for $X \in \mathcal{X}'$ **do**
    if $X \in \mathcal{C}$ **then**
      $\pi^* = \arg\max\{\pi : \hat{I}(\mathcal{G}, X_\pi; Y), \pi \in \Pi_{l,n}(Q)\}$
      $\mathcal{G} = \mathcal{G} \cup \{X_{\pi^*}\}$
    else
      $\mathcal{G} = \mathcal{G} \cup \{X\}$
  return $\hat{I}(\mathcal{G}; Y)$

strategy Q, consistent discrete estimator $\hat{I}$, and maximum number of bins l, the estimation process starts by marginally sorting the $X \in \mathcal{X}$ according to $Q, l, \hat{I}$ (Q, l, are used for $X \in \mathcal{C}$), and create the empty set $\mathcal{G}$ for discrete variables. Then, continuous variables $X \in \mathcal{C}$ are discretized jointly with $\mathcal{G}$ and added to $\mathcal{G}$, while the discrete $X \in \mathcal{D}$ are added to $\mathcal{G}$. The mixed estimator result is then $\hat{I}(\mathcal{G}; Y)$. If T_Q is the cost for optimization based on Q, $T_{\hat{I}}$ the cost of estimator $\hat{I}$, and $|\mathcal{C}| = m$, the algorithm complexity is dominated by $O(mT_QT_{\hat{I}})$. For the remainder, we refer to a specific instantiation of the mixed estimator with the estimator and partitioning technique choices, e.g., $\hat{I}_{\text{pl}}$ with EF.

5.3.2 Statistical efficiency

Now that an optimization framework is established, we shift our attention to a brief discussion regarding appropriate qualities discrete consistent estimators should possess for the task of FDD.

We are mainly after estimators that allow for efficient discovery, i.e., come with the means for high-dimensional exhaustive and heuristic search. An optional, yet important requirement, is admitting optimal substructure for applying DP and giving access to a large set of candidate partitions in polynomial time. The

third dimension is that of statistical efficiency: the estimator should give robust estimation from limited data samples for both the partitioning process, as well as the discovery process. Let us mainly focus on the last requirement, and demonstrate how the consistency of an estimator, alone, does not satisfy it. As an example, we consider the plugin estimator $\hat{I}_{\text{pl}}$ and start with the following lemma.

Lemma 5.3.1. *Given continuous variable C, discrete set $\mathcal{G}$, discrete target Y, and maximum number of bins l, we have that*

a) $\hat{I}_{pl}(\mathcal{G}, C_\pi; Y) \leq \hat{I}_{pl}(\mathcal{G}, C_{\pi'}; Y)$, *for all* $\pi, \pi' \subseteq \Pi_{l,n}$ *with* $\pi \preceq \pi'$

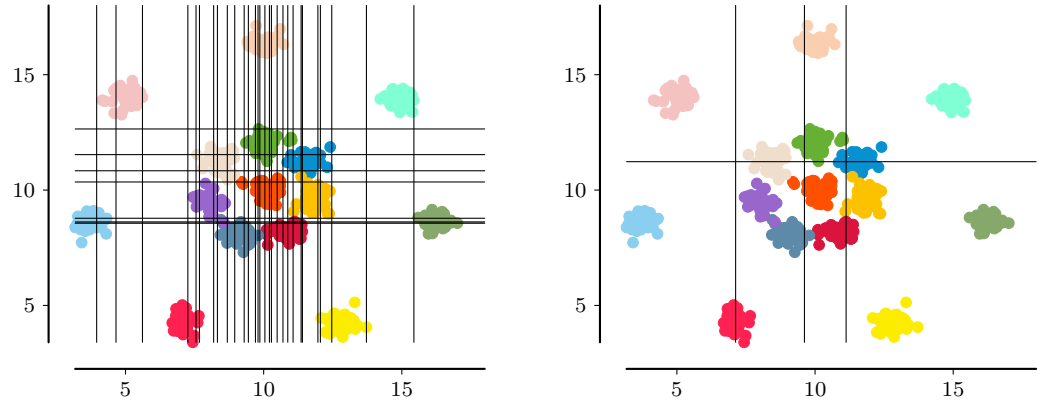
b) $\hat{I}_{pl}(\mathcal{G}, C_{Q_k^{\text{EF}}}; Y) \leq \hat{I}_{pl}(\mathcal{G}, C_{Q_{2k}^{\text{EF}}}; Y)$ *for* $k = 1, \ldots, \lfloor l/2 \rfloor$

c) $\hat{I}_{pl}(\mathcal{G}, C_\pi; Y) \leq \hat{I}_{pl}(\mathcal{G}, C_{Q_l^{\text{EF}}}; Y)$, *for all* $\pi \in \Pi_{l,n}(Q^{\text{cOP}})$

Proof. Recall the specialization relation (Def. 3.1.1): for two discrete variables A, B, we say that B is a specialization of A, denoted as $A \preceq B$, if for all $i, j \in [n]$ with $A(i) \neq A(j)$, it holds $B(i) \neq B(j)$. It is clear that a refinement relation for $\pi \preceq \pi'$, corresponds to a specialization relation for $C_\pi \preceq C_{\pi'}$. Finally, we have that for three variables A, B, C with $A \preceq B$, that $\hat{I}_{\text{pl}}(A; C) \leq \hat{I}_{\text{pl}}(B; C)$ (Prop. 3.1.1).

For (a), we have that $C_\pi \preceq C_{\pi'}$ for any $\pi \preceq \pi'$, and hence $\hat{I}_{\text{pl}}(\mathcal{G}, C_\pi; Y) \leq \hat{I}_{\text{pl}}(\mathcal{G}, C_{\pi'}; Y)$. For (b) and (c), we have that $Q_k^{\text{EF}} \preceq Q_{2k}^{\text{EF}}$ for $k = 1, \ldots, \lfloor l/2 \rfloor$, and $\pi \preceq Q_l^{\text{EF}}$ for all $\pi \in \Pi_{l,n}(Q^{\text{cOP}})$, respectively. The two statements then follow from (a). □

Lemma 5.3.1 states that $\hat{I}_{\text{pl}}$ considers refinements to be at least as good of a choice. However, unlike the quantization process in the population, refined partitions on a sample do not necessarily lead to a better estimation error, but rather to overfitting. For $\hat{I}_{\text{pl}}$ in particular, the overfitting is controlled by the statistical bias that is a function of the domain sizes $S_{\mathcal{G},C_\pi}$ and S_Y [Rou99]. In a nutshell, larger $|\pi|$ implies more bias for $\hat{I}_{\text{pl}}(\mathcal{G}, C_\pi; Y)$, and hence $\hat{I}_{\text{pl}}$ tends to trivially select the most refined partition for a Q and l. We demonstrate this behavior with the following example.



(a) $\hat{I}_{\text{pl}}, l = 30, c = 3$ **(b)** $\hat{I}_{\text{MIC}}, c = 3$

Figure 5.2: Resulting partitions from estimating mutual information on a clustering dataset in $\mathbb{R}^2$. Plugin $\hat{I}_{\text{pl}}$ combined with two versions of OP. The target Y is the cluster assignment (colored). (Example 5.3.1)

Example 5.3.1. *In this example we investigate the resulting partitions from estimating mutual information on a clustering dataset in $\mathbb{R}^2$, where the target variable Y is the cluster assignment. The dataset has 600 data points and 15 clusters [FS18], and we use C_1, C_2, to refer to x and y-axis, respectively. We are after an estimate of $\hat{I}_{pl}(C_1, C_2; Y)$, and consider two versions of* cOP*. For the first, we use a fixed $l = 30$ for both C_i, while for the second we use $l = g(n, b, \mathcal{G}, Y)$ per C_i that is proposed in [RRF+11], where $g(n, b, \mathcal{G}, Y) = \lceil n^b / (\prod_{G \in \mathcal{G}} V_G V_Y) \rceil$. We set $b = 0.6$ that is suggested by the authors, and refer to this estimator as the* ***MIC estimator*** *$\hat{I}_{\text{MIC}}$. For both we use $c = 3$. We present the results in Figure 5.2. On the left, we observe that $\hat{I}_{pl}$ indeed selects for C_1 the most refined partition possible, i.e., Q_{30}^{EF}, as the lemma suggests. For C_2, there are 8 bins, but only because there is a perfect cluster separation already for a total of 240 bins in $\mathbb{R}^2$. On the right, $\hat{I}_{\text{MIC}}$ has a maximum $l = 4$ for C_1, and $l = 2$ for C_2 (for C_2, $\mathcal{G}$ already contains the discrete C_1). Again, $\hat{I}_{pl}$ selects the maximum number of bins for both variables, but here we actually observe underfitting caused by the criterion $l = g(n, b, \mathcal{G}, Y)$.*

We see that $\hat{I}_{\text{pl}}$ can easily under/over fit the data during the partition process, even with more elaborate criteria for l, e.g., the $g(n, b, \mathcal{G}, Y)$ used in MIC [RRF+11]. Note that $\hat{I}_{\text{MIC}}$ is an inherent part of MIC, as it identifies the best partition for each

$k = 1, \ldots, l = g(n, \emptyset, Y)$. The best partitions are afterwards penalized by their size k, which is not a statistical adjustment accounting for the biased estimates. It is demonstrated that MIC overfits on noisy data [KA14].

In addition to the partition process, we consider the task of FDD, i.e., finding the $\mathcal{X}^* \subseteq \mathcal{I}$ maximizing $F(\mathcal{X}^*; Y)$. Translating this to our example, it would mean to identify the top clustered data out of a potentially huge candidate space of varying dimensionalities. For FDD, the $\hat{F}_{\text{pl}}$ fails by trivially considering $\mathcal{X}^* = \mathcal{I}$ to be a maximizer. As we see, choosing an estimator for FDD is non-trivial: besides being "optimizable" for efficient algorithms and exhibiting optimal substructure, estimators need to be statistically efficient and robust against choices for l, Q, and varying dimensionalities. In Section. 5.5 we evaluate different choices for $\hat{I}$ and Q.

5.4 Robust functional dependency discovery from mixed data

In this section, we show the permutation mutual information estimator $\hat{I}_0$ exhibits optimal substructure and then give algorithms for Eq.(3.1) and mixed data.

Theorem 5.4.1. *Given discrete variables $\mathcal{G}$, continuous X, discrete Y, and maximum number of bins l, the optimization problem*

$$\max_{\pi \in \Pi_{l,n}(Q^{\text{cOP}})} \hat{I}_0(\mathcal{G}, X_\pi; Y)$$

exhibits for $1 < l \leq m \leq n$ the optimal substructure

$$f(l, m) = \max_{1 \leq i < m} \{ \frac{i}{m} f(l-1, i) + \frac{m-i}{m} \hat{I}_0(G \,; Y \mid i+1, m) \} \ ,$$

with

$$f(l, m) = \max_{\pi \in \Pi_{l,m}} \hat{I}_0(\mathcal{G}, X_\pi; Y \mid 1, m) \ ,$$

and

$$\hat{I}_0(. \,; Y \mid u, v) = \hat{I}_{pl}(. \,; Y \mid u, v) - \sum_{\sigma \in S_n} \hat{I}_{pl}(. \,; Y_\sigma \mid u, v)/n! \ ,$$

where $\hat{I}_{pl}(.\,;.\,|\,u,v)$ *with* $u,v \in [n], v \geq u$ *is the empirical mutual information restricted to data samples* $\{i \in [n]\,|\,X_s(u) \leq X(i) \leq X_s(v)\}$.

Proof. Let us assume w.l.o.g. that $f(m,l)$ corresponds to partition $\pi^* = \{S_1, \ldots, S_l\}$ of l bins, and $V_{X_{\pi^*}} = \{x_1, \ldots, x_l\}$. We use $c_j = \sum_{i=1}^{j} n_{x_i}$ for $j \in [l]$, and n^σ denotes the empirical count after a permutation $\sigma \in S_n$ for Y. We have $f(l,m) =$

$$
\begin{aligned}
&- \frac{1}{n!} \sum_{\sigma\in S_n} \sum_{y\in Y} \sum_{g\in\mathcal{G}} \sum_{i=1}^{l} \frac{n^\sigma_{ygx_i}}{m} \log \frac{n^\sigma_{ygx_i}}{n_{gx_i}} + \sum_{y\in Y} \sum_{g\in\mathcal{G}} \sum_{i=1}^{l} \frac{n_{ygx_i}}{m} \log \frac{n_{ygx_i}}{n_{gx_i}} \\
=& - \frac{1}{n!} \sum_{\sigma\in S_n} \sum_{y\in Y} \sum_{g\in\mathcal{G}} \sum_{i=1}^{l-1} \frac{n^\sigma_{ygx_i}}{m} \log \frac{n^\sigma_{ygx_i}}{n_{gx_i}} - \frac{1}{n!} \sum_{\sigma\in S_n} \sum_{y\in Y} \sum_{g\in\mathcal{G}} \frac{n^\sigma_{ygx_l}}{m} \log \frac{n^\sigma_{ygx_l}}{n_{gx_l}} \\
&+ \sum_{y\in Y} \sum_{g\in\mathcal{G}} \sum_{i=1}^{l-1} \frac{n_{ygx_i}}{m} \log \frac{n_{ygx_i}}{n_{gx_i}} + \sum_{y\in Y} \sum_{g\in\mathcal{G}} \frac{n_{ygx_l}}{m} \log \frac{n_{ygx_l}}{n_{gx_l}} \\
=& - \frac{c_{l-1}}{mn!} \sum_{\sigma\in S_n} \sum_{y\in Y} \sum_{g\in\mathcal{G}} \sum_{i=1}^{l-1} \frac{n^\sigma_{ygx_i}}{c_{l-1}} \log \frac{n^\sigma_{ygx_i}}{n_{gx_i}} - \frac{m-c_{l-1}}{mn!} \sum_{\sigma\in S_n} \sum_{y\in Y} \sum_{g\in\mathcal{G}} \frac{n^\sigma_{ygx_l}}{m-c_{l-1}} \log \frac{n^\sigma_{ygx_l}}{n_{gx_l}} \\
&+ \frac{c_{l-1}}{mn!} \sum_{y\in Y} \sum_{g\in\mathcal{G}} \sum_{i=1}^{l-1} \frac{n_{ygx_i}}{c_{l-1}} \log \frac{n_{ygx_i}}{n_{gx_i}} + \frac{m-c_{l-1}}{mn!} \sum_{y\in Y} \sum_{g\in\mathcal{G}} \frac{n_{ygx_l}}{m-c_{l-1}} \log \frac{n_{ygx_l}}{n_{gx_l}} \\
=& - \frac{c_{l-1}}{m} \Big(\frac{1}{n} \sum_{\sigma\in S_n} \sum_{y\in Y} \sum_{g\in\mathcal{G}} \sum_{i=1}^{l-1} \frac{n^\sigma_{ygx_i}}{c_{l-1}} \log \frac{n^\sigma_{ygx_i}}{n_{gx_i}} - \sum_{y\in Y} \sum_{g\in\mathcal{G}} \sum_{i=1}^{l-1} \frac{n_{ygx_i}}{c_{l-1}} \log \frac{n_{ygx_i}}{n_{gx_i}} \Big) \\
&- \frac{m-c_{l-1}}{mn!} \Big(\sum_{\sigma\in S_n} \sum_{y\in Y} \sum_{g\in\mathcal{G}} \frac{n^\sigma_{ygx_l}}{m-c_{l-1}} \log \frac{n^\sigma_{ygx_l}}{n_{gx_l}} - \sum_{y\in Y} \sum_{g\in\mathcal{G}} \frac{n_{ygx_l}}{m-c_{l-1}} \log \frac{n_{ygx_l}}{n_{gx_l}} \Big) \\
=& \frac{c_{l-1}}{m} \hat{I}_0(G, X_{\pi^*\setminus\{S_l\}}; Y \,|\, 1, c_{l-1}) + \frac{m-c_{l-1}}{m} \hat{I}_0(G\,; Y \,|\, c_{l-1}+1, m) \\
=& \frac{c_{l-1}}{m} f(l-1, c_{l-1}) + \frac{m-c_{l-1}}{m} \hat{I}_0(G\,; Y \,|\, c_{l-1}+1, m) \ ,
\end{aligned}
$$

where the last equality holds, otherwise we could increase $f(l,m)$ with a different partition for the first c_{l-1} points. Hence, for $l, m > 1$ we arrive at the following optimal substructure recursive relation

$$
f(l,m) = \max_{1\leq i<m} \{ \frac{i}{m} f(l-1, i) + \frac{m-i}{m} \hat{I}_0(G\,; Y \,|\, i+1, m) \} \ .
$$

□

Now that optimal substructure has been established, we shift our attention to optimization algorithms for FDD. For mixed data, first note the problem of maximizing $\hat{F}_0$ remains NP-hard. Second, the admissible bounding functions $\bar{f}_{\text{mon}}(\mathcal{X}, Y)$ and $\bar{f}_{\mathbf{spc}}(\mathcal{X}, Y)$ (Sec. 3.3) remain admissible as they are independent of "future" partitions for $\mathcal{X}'$ with $\mathcal{X} \subseteq \mathcal{X}'$. However, unlike the discrete case, here evaluating $\hat{I}_0(\mathcal{X}; Y)$ for a candidate $\mathcal{X} = \{\mathcal{D}, \mathcal{C}\}$ is more expensive—Algorithm 5 sorts in decreasing order of marginal $\hat{I}_0$, and performs $|\mathcal{C}|$ discretizations. For efficiency, we first remove the repetitive sorting by sorting $\mathcal{I}$ initially instead. Then the alphabetic refinement operator only refines with variables of smaller marginal mutual information. Second, we apply the following heuristic: once a variable $C \in \mathcal{C}$ has been discretized, it remains discretized for the remaining of the search branch. With these, we then adapt and use Algorithms 2 and 3 from Section 3.4. It is important to note that since we sort $\mathcal{I}$ initially, the admissible heuristic of Webb [Web95] to assign the most refinement operators to the least promising nodes (i.e., smallest potential) is not applicable here as it violates the ordering.

5.5 Evaluation

In this section, we perform an evaluation on the different aspects of our FDD solution for mixed data. In particular, we investigate the statistical performance of various estimators coupled with partitioning techniques on synthetic data, we evaluate the proposed discovery algorithms on real-world benchmark data, and finally, we qualitatively analyze the partitions selected from estimation.

5.5.1 Estimator performance

First, we focus on the statistical performance of mixed estimator configurations. We are interested in their consistency with regards to the FDD process. For this, we generate data from models governing functional relationships for which we know the population values for mutual information, perform FDD with exhaustive

search to obtain the estimated value of the maximizer variable set, and then plot curves corresponding to absolute estimation error.

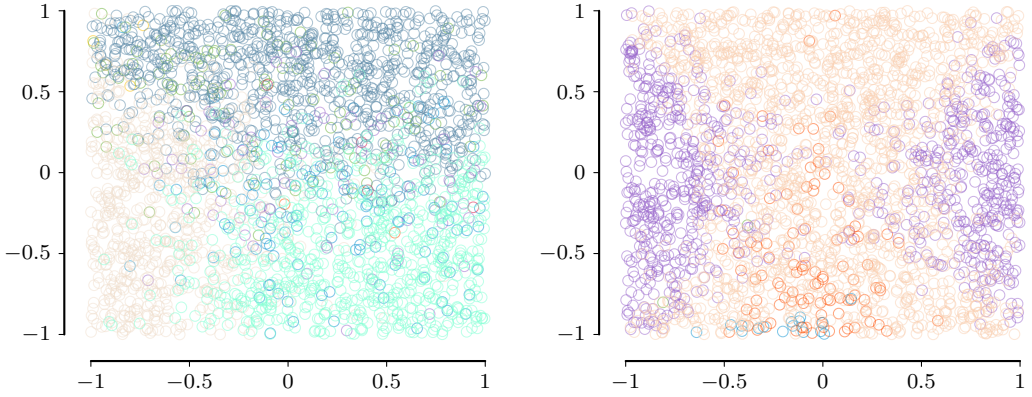
In this experiment, we model our functional relationships with the class of **generalized linear models**. We consider a set of four continuous random variables $\mathcal{I} = \{X_1, X_2, X_3, X_4\}$, and one categorical variable Y, and distinguish two cases of functional relationship: $\mathbb{E}[Y \mid \mathcal{I}] = f^{-1}(\alpha_0 + \sum_{j=1}^{4} \alpha_j X_j)$ and $\mathbb{E}[Y \mid \mathcal{I}] = f^{-1}(\beta_0 + \sum_{i=1}^{3} \beta_i \sum_{j=1} \alpha_{j,i} g_i(X_j))$, where f is an appropriate link function and $g_1(X) = \log(X+2), g_2(X) = X^2, g_3(X) = \cos(2X)$ are non-linear variable transformations. We use $h \in \{\text{lin}, \text{nlin}\}$ to indicate the former and latter cases respectively. The coefficients α, β, follow a bimodal Gaussian distribution that uniformly selects one of $\mathcal{N}(-\log(10), 1)$ and $\mathcal{N}(\log(10), 1)$. The means log(10) and -log(10) are chosen such that the respective classes for binary Y (positive for log(10) and negative for $-\log(10)$), are 10 times more likely. To cover a wider range of scenarios, we further parametrize the models in two ways: we consider a varying number $e \in \{1, 2, 3\}$ of explanatory variables, with the remaining $4 - e$ receiving weights $\alpha = 0$, and we use three domain sizes $d \in \{2, 5, 10\}$ for Y.

For our **generative models** $p_{\alpha,\beta}(\mathcal{I}, Y)$, variables X_j follow a uniform $\mathbf{U}(-1, 1)$ and Y a multinomial with expectations as above. We omit α, β from notation for readability. Given parameters $d \in \{2, 5, 10\}$, $e \in \{1, 2, 3\}$, and $h \in \{\text{lin}, \text{nlin}\}$, we denote the resulting models with $p^l_{e,d}(\mathcal{I}, Y)$. For the conditional $p^h_{e,2}(Y \mid \mathcal{I})$ we use the sigmoid function (i.e., logit link function), and the softmax for $p^h_{e,\{5,10\}}(Y \mid \mathcal{I})$ (i.e., multinomial logit). The analytic expressions are found in Table 5.1. With these, for any set of coefficients α, β, we can compute the population value $I(\mathcal{I}; Y)$. To sample data from the models, we first randomly and uniformly sample 90 conditional probability distributions $p^{(i)}, i = 1, \ldots, 90$, 5 for each combination of e, d, h. To make the results comparable, we ensure for each $p^{(i)}$ the population value $F(p^{(i)})$ lies in $(0, 0.5]$. We denote with $\mathcal{P}^l_{e,d}$ the sets of $p^{(i)}$ corresponding to specific e, d, l. For example, $\mathcal{P}^{\text{lin}}_{2,2}$ is the set of the 5 $p^{(i)}$ corresponding to $d = 2, e = 2, h = \text{lin}$. We consider data sizes $n \in \{20, 40, 80, 160, 320, 640, 1280, 2560\}$, and for each $p^{(i)}$ and n, we sample 50 datasets $\mathbf{D}^{(i)}_{n,j}, j \in [1, 50]$. Two sampled datasets are illustrated in Figure 5.3.

Now, given these data, we perform the FDD task with input variables $\mathcal{I}$ and

Table 5.1: Analytic expressions for $p^h_{e,d}(Y \,|\, \mathcal{I})$

parameters	analytic expressions
$h = \text{lin}, d = 2, Y = 1$	$\frac{1}{1+e^{-(\alpha_0+\sum_{j=1}^{4}\alpha_j X_j)}}$
$h = \text{nlin}, d = 2, Y = 1$	$\frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{3}\beta_i \sum_{j=1}^{4}\alpha_{j,i} g_i(X_j)}}$
$h = \text{lin}, d \in \{5, 10\}, Y = q$	$\frac{e^{\alpha_{0,q}+\sum_{j=1}^{4}\alpha_{j,q}X_j}}{\sum_{z=1}^{d} e^{\alpha_{0,z}+\sum_{j=1}^{4}\alpha_{j,z}X_j}}$
$h = \text{nlin}, d \in \{5, 10\}, Y = q$	$\frac{e^{\beta_{0,q}+\sum_{i=1}^{3}\beta_{i,q}\sum_{j=1}^{4}\alpha_{j,i}g_i(X_j)}}{\sum_{z=1}^{d} e^{\beta_{0,z}+\sum_{i=1}^{3}\beta_{i,z}\sum_{j=1}^{4}\alpha_{j,i}g_i(X_j)}}$



(a) linear **(b)** non-linear

Figure 5.3: Example data sampled from generalized linear models. Here we have 10 classes (colored), 2 explanatory variables, and $n = 2560$ data points

target Y, considering different estimator/partitioning configurations combined with exhaustive search. We consider the plugin $\hat{I}_{\text{pl}}$, the permutation $\hat{I}_0$, the MIC $\hat{I}_{\text{MIC}}$ (Example 5.3.1), and the chi-square estimator $\hat{I}_{\chi,\alpha}$ (Sec. 3.5.1). To evaluate the performance, we use the **absolute estimation error** tailored for FDD, defined as $r_n(\hat{F}_{mx}, p^{(i)}) = \mathbb{E}[|F(p^{(i)}) - \hat{F}_{mx}(\mathcal{X}^*_{i,j,n}; Y)|]$, where $F(p^{(i)})$ is the population fraction of information value for model $p^{(i)}$, and $\mathcal{X}^*_{i,j,n} \subseteq \mathcal{I}$ is the maximizer on $\mathbf{D}^{(i)}_{n,j}$ for a configuration $\hat{F}_{mx}$. We use the fraction of information instead in order to have the error in $[0, 1]$. The expected value is with respect to $j \in [1, 50]$. We average the absolute errors across different $p^{(i)}$ to obtain averages of the form $r_n(\hat{F}_{mx}, \mathcal{P}^{\{\text{lin,hlin}\}}_{[a,b],\{2,5,10\}})$. For example, $r_n(\hat{F}_{mx}, \mathcal{P}^{\{\text{lin,hlin}\}}_{[1,3],\{2,5,10\}})$ corresponds to the average

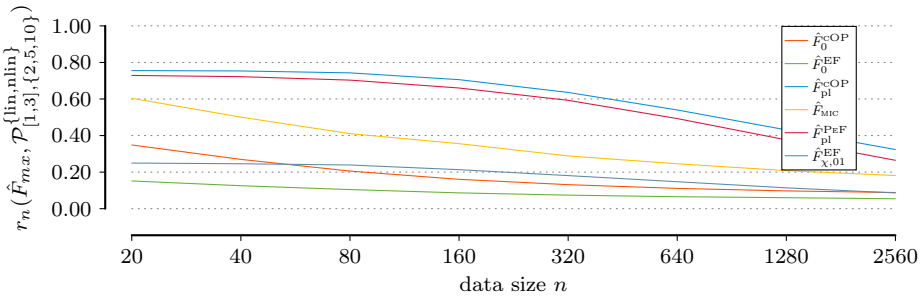


Figure 5.4: Absolute estimation error averaged across all models $p^{(i)}$.

absolute error across all 90 models $p^{(i)}$, while the $r_n(\hat{F}_{mx}, \mathcal{P}^{\text{hlin}}_{[1,3],2})$ would be the average for $p^{(i)} \in \mathcal{P}^{\text{hlin}}_{1,2} \cup \mathcal{P}^{\text{hlin}}_{2,2} \cup \mathcal{P}^{\text{hlin}}_{3,2}$.

We start with Figure 5.4 and plot the average error curves across all $p^{(i)}$, for $\hat{F}_0$ with cOP and EF, $\hat{F}_{\text{pl}}$ with cOP, $\hat{F}_{\text{MIC}}$, and $\hat{F}_{\chi,\alpha}$ with EF. For $\hat{F}_{\chi,\alpha}$, we tested both $\alpha = 0.95$ and 0.99, and show the latter that has better performance. For cOP and EF we use maximum number of bins $l = 5$, and for cOP $c = 2$. In addition, we consider $\hat{F}_{\text{pl}}$ with pre-discretized data in 5 equal-frequency bins as a baseline, which we refer to as PEF. Let us focus first on the three uncorrected configurations, i.e., $\hat{I}_{\text{pl}}$ with cOP, PEF, and $\hat{F}_{\text{MIC}}$. All under-perform, with the highest errors and slower convergence rates. Interestingly, we see that PEF performs better than cOP, despite both having $l = 5$. This behavior is attributed to the $\mathcal{I}$ being uniform and independent, and while PEF is well-suited for this, $\hat{F}_{\text{pl}}$ with cOP overfits by finding joint effects. For $\hat{F}_{\text{MIC}}$, the convergence is better, but only because the maximum number $l = g(n, 0.6, \mathcal{G}, Y)$ decreases per $X \in \mathcal{X}$, and the subsequent "coarser" X exchange overfitting for underfitting (as in Example 5.3.1). Moving on to the two corrected estimators $\hat{F}_0, \hat{F}_{\chi,01}$, combined with EF, we observe lower errors and faster convergence, with $\hat{F}_0$ showing the best performance. Lastly, the permutation $\hat{F}_0$ combined with cOP has higher error for smaller number of samples, but performs well in terms of convergence speed and "catches" up. Note that $X \in \mathcal{I}$ are uniform, and EF meets this requirement. The cOP with $c = 5, c = 2$, considers only one equal-frequency partition, that of Q_5^{EF}, which cannot be supported for small n due to the correction.

Now let us briefly focus on averages over different configurations. Note that all

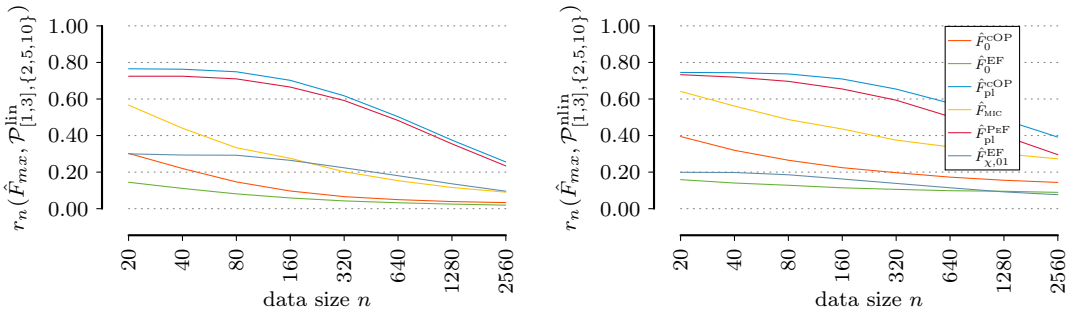


Figure 5.5: Absolute estimation error averaged across all models $p^{(i)}$ with (right) and without (left) the non-linear layer.

$p^{(i)}, i \in [90]$, are different models and for the following figures, one should focus mainly on the convergence speed comparison between two plots. In Figure 5.5 we show the results averaged over the 45 $p^{(i)}$ corresponding to the additional non-linear layer in the functional relationship, i.e., $h = \text{nlin}$ (right), and over the 45 $p^{(i)}$ with $h = \text{lin}$ (left). Between the two, we observe that convergence speeds are better for the case $h = \text{lin}$, as expected. We also see that EF performs well in both cases. Moving on, we average over the 30 $p^{(i)}$ where there is only one explanatory variable, i.e., $e = 1$, and the 30 $p^{(i)}$ with $e = 3$. Additionally, we average over the 30 $p^{(i)}$ with target domain size $V_Y = 2$, i.e., $d = 2$, and the 30 $p^{(i)}$ with $d = 10$. In Figure 5.6 and 5.7 corresponding to the former and latter, we observe that methods are robust against number of explanatory variables, but there is over fitting for $S_Y = 2$. Here, the dependency is "easier" to infer and estimators may select supersets $\mathcal{X}'$ of $\mathcal{X}^*$ that have the same value $F(\mathcal{X}^*; Y) = F(\mathcal{X}'; Y)$, but on the sample there is overestimation. Note that in Figure 5.7 we do not plot $\hat{F}_{\text{MIC}}$ as it could not terminate due to the scale of the experiment, since for $S_Y = 2$ there is a large number of candidate partitions to consider.

Overall, the permutation $\hat{F}_0$ shows the best performance with EF that accurately fits the uniform data $\mathcal{I}$. Combined with cOP that for $l = 5, c = 2$, considers mostly non-uniform partitions, the error is higher for small n, but the convergence is fast. The EF should be preferred when assumptions are met, i.e., uniformity, and cOP for exploratory scenarios.

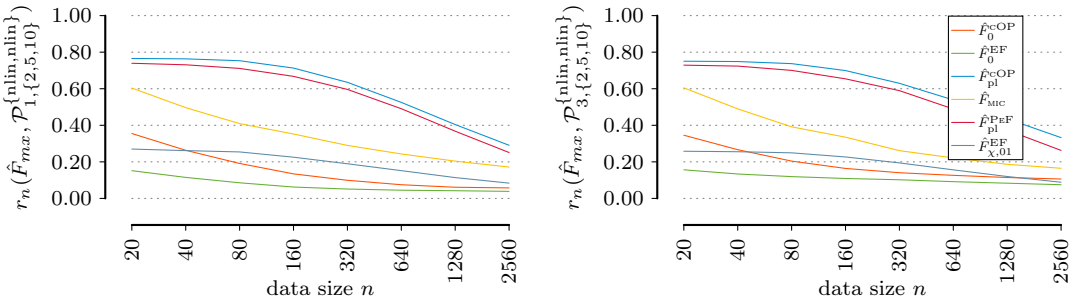


Figure 5.6: Absolute estimation error averaged across all models $p^{(i)}$ with one (left) and three (right) explanatory variables.

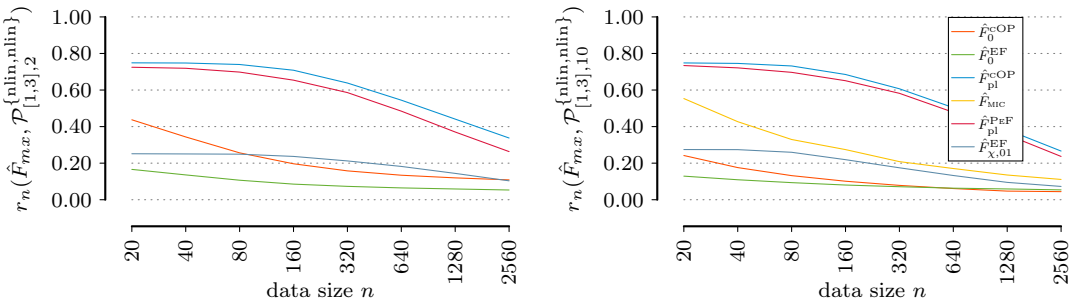


Figure 5.7: Absolute estimation error averaged across all models $p^{(i)}$ with target domain size $V_Y = 2$ (left), and 10 (right).

5.5.2 Optimization performance

Here, we perform FDD on data from the KEEL data repository [SRAFFH+11]. In particular, we use all classification datasets with mixed and continuous input attributes $\mathcal{I}$ and no missing values, resulting in 29 datasets with 25000 and 30 rows and columns on average, respectively. There are 12, 6, and 2, continuous, ordinal discrete, and nominal attributes, respectively, on average per dataset, all summarized in Table 5.2. We employ the two algorithms Opus and Grd (Sec. 3.4) with the chain bounding function $\bar{f}_{\text{chn}}$ (Sec. 3.3) to retrieve the top solution $\mathcal{X}^* \subseteq \mathcal{I}$, both combined with EF and cOP for $l = 5, c = 2$. To increase the difficulty, the ordinal discrete $D \in \mathcal{I}$ per dataset are also partitioned when $V_D \geq l$. For Opus, and for each of EF and cOP, we set α to be the highest possible in increments of 0.05, such that they terminate in less than 1 hour. For Opus, we report in

Table 5.2 the α values, the runtime, the size $|\mathcal{X}^*|$ of the solution, and the value $\hat{F}_0(\mathcal{X}^*; Y)$. Similarly for GRD, we report runtime and $\hat{F}_0(\mathcal{X}^*; Y)$. The runtimes are averaged over 3 independent executions. This experiment is executed on an Intel Xeon E5-2643 v3 with 256 GB memory. Our code is online for research purposes.[1]

We start with OPUS and α values. For both EF and COP, the average α value for OPUS to complete in ≤ 1 hour is 0.81. There are 14 datasets for EF and 15 for COP with $\alpha = 1$, which corresponds to an exact solution, while there are 6 datasets for both with $\alpha \in [0.8, 1)$. Here, we see that both methods offer good guarantees with a budget of 1 hour. Regarding the cardinality $|\mathcal{X}^*|$ of the solutions, for EF they have size 3.8 on average, while for COP 3.5. Again, the two partitioning techniques show similar performance, with COP returning slightly smaller sets. We hypothesize this is due to the ability of COP to better adapt on data, and hence extract more information with fewer attributes. Time-wise, EF and COP require 599 and 743 seconds on average. The COP is slower, as expected. Finally, the average quality of the solution is 0.52 and 0.53 for EF and COP, respectively, with COP recovering 1% more information by considering more candidate partitions. The greedy algorithm is efficient, with EF requiring 51 seconds on average and COP 43. Interestingly, the quality of the solutions are higher than OPUS, with 0.53 and 0.55 on average for EF and COP, respectively. In fact, the solutions of GRD have roughly the same quality as those of OPUS for high α values, while for smaller α GRD has better quality.

Overall, we observe that both algorithms OPUS and GRD, with both partitioning techniques EF and COP, are very effective in practice. For truly exploratory scenarios, COP should be preferable over EF, unless the assumptions on the data distributions are met by EF. The branch-and-bound algorithm should be used whenever solution guarantees are required. The greedy algorithm, however, is very efficient and hence a better candidate for larger datasets. In addition, it shows good performance in terms of solution quality.

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

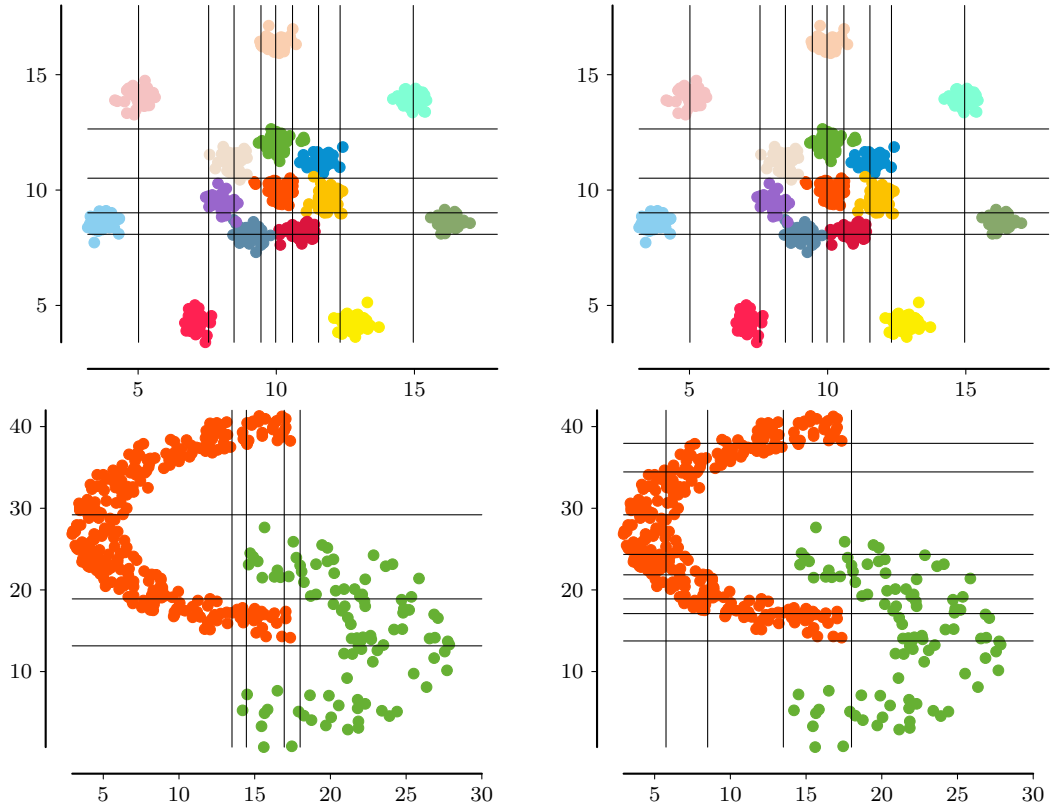
5.5.3 Qualitative analysis

Here, we present the resulting partitions from estimating mutual information on clustering datasets in $\mathbb{R}^2$, where the target Y is the cluster assignment [FS18]. We denote the variables corresponding to x and y-axis with C_1 and C_2, respectively. We use the permutation $\hat{I}_0$ with optimal (cOP) and equal frequency (EF) partitioning. For both cOP and EF we set the maximum number of bins l to 10, and use $c = 3$ in order to have 30 initial equal-frequency bins for cOP. This allows to investigate whether $\hat{I}_0$ overfits by having access to more candidate partitions. For all methods, C_1 is discretized first for better comparison. We present the results in Figure 5.8.

The first dataset has 15 clusters. The $\hat{I}_0$ with both cOP and EF results in the same partition in 40 bins. Here, $\hat{I}_0$ performs well at separating the clusters. The second dataset has 2 clusters. The $\hat{I}_0$, cOP, configuration with 20 bins in $\mathbb{R}^2$ perfectly separates the clusters, while with EF there are 45 bins. In addition to these, we used $\hat{I}_{\chi,01}$ with EF, $\hat{I}_{\text{pl}}$ with cOP, and $\hat{I}_{\text{MIC}}$. We report that $\hat{I}_{\chi,01}$ has identical results with $\hat{I}_0$ and EF, $\hat{I}_{\text{pl}}$ partitions with the maximum number of bins, while $\hat{I}_{\text{MIC}}$ selects an overly refined partition for C_1, and mostly 2 bins for C_2. Overall, we see that $\hat{I}_0$ results in good partitions for both EF and cOP. For the latter in particular, there is better class separation with less bins. This indicates that $\hat{I}_0$ with cOP selects good partitions, without overfitting on larger spaces of candidates, and can better adapt on more "exotic" distributions. For EF, both $\hat{I}_0$ and $\hat{I}_{\chi,01}$ select finer-grained partitions. The $\hat{I}_{\text{pl}}$ with cOP and $\hat{I}_{\text{MIC}}$ under-perform, as expected.

5.6 Discussion and conclusions

We considered the task of robust functional dependency discovery from mixed data. We proposed a mixed mutual information estimator framework based on the theoretical process of random variable quantization. We demonstrated how it can be applied for the task of functional dependency discovery from empirical data and instantiated it with the permutation fraction of information. Lastly, we gave algorithms for exact, approximate, and heuristic search. The experimental evaluation showed that the estimator has desired statistical properties, the bounding



(a) $\hat{I}_0^{\text{cOP}}, l = 10, c = 3$ **(b)** $\hat{I}_0^{\text{EF}}, l = 10$

Figure 5.8: Resulting partitions from estimating mutual information on clustering datasets in $\mathbb{R}^2$. The target Y is the cluster assignment (colored).

functions remain very effective with both exhaustive and heuristic algorithms, and the greedy algorithm provides again solutions that are nearly optimal. A case study with data from Materials Science (Fig. 5.1), as well as qualitative experiments regarding the partitioning process, indicate that our proposed framework indeed discovers informative dependencies.

5.6.1 Maximum number of partitions l

The various sub-linear to n criteria discussed in this chapter, e.g. $\log_2(n)$, correspond to methods that consider univariate pairs. On the one hand, naively extending these for each $C \in \mathcal{C}$ can lead to an exponential increase of partitions in

the $|\mathcal{C}|$-dimensional space with each data point falling in one hypercube, violating therefore consistency even for optimal estimators (Thm. 5.2.2). For example, let us assume $n = 10000$. We have $l = \log_2(n) \approx 13$, and we can already for $|\mathcal{C}| = 4$ arrive at one point per hypercube. On the other hand, a more appropriate way would be to set $\log_2(n)$ as the maximum number of hypercubes allowed in $\mathbb{R}^{|\mathcal{C}|}$, but this can be very conservative–in our example, it would mean to place 10000 data points in 13 hypercubes, regardless of $|\mathcal{C}|$. Note that these calculations are done independently of $\mathcal{D}$ that only exacerbates this behavior. For our purposes, we instead considered a fixed l, e.g., 5. This way, and combined with a corrected estimator, we better control for the aforementioned problems. While one could potentially derive a joint criterion accounting for both $|\mathcal{X}|$ and n, we did not consider this investigation here.

5.6.2 Unsupervised scenario

In this chapter we considered the supervised scenario. The unsupervised case investigated in Chapter 4, i.e., maximizing the normalized total correlation on discrete data, poses many challenges. In particular, the search space is enumerated under a strict partial order that depends on the marginal entropies of the discrete variables. For continuous attributes, their discrete marginal entropies are not known in advance since they get discretized during optimization. While the possibility for greedy search remains, it is not trivial to design pruning rules to be used for exhaustive search. A possible direction would be to consider equal-frequency discretization and ensure that refinements use as upper-bound l the domain size of the previous attribute.

5.6.3 Future work

For future work, it would be interesting to consider experiments with generalized linear models and correlated explanatory variables, e.g., Gaussian with non-diagonal covariance matrix. That would highlight the importance of the joint discretization our framework considers. Moreover, adaptive partitioning could be applicable, which would allow to consider a different class of candidate partitions

and potentially more data efficient. Another direction is establishing the regularized supervised discretization process and add it to a large family of supervised discretization techniques such as the MDL approach by Fayyad and Irani [FI93].

Table 5.2: Datasets used in Section 5.5.2. Number of attributes is subdivided into real, integer, and nominal. With / we separate the results for EF (left) and cOP (right). The α column corresponds to the highest possible α value such that Opus terminates ≤ 1 hour. The cardinality of the solution is column $|\mathcal{X}^*|$. Last two columns is the score for $\mathcal{X}^*$ by Opus and Grd, respectively.

						time(s)		$\hat{F}_0(\mathcal{X}^*;Y)$	
dataset	#attr. (r/i/n)	#rows	#clas.	α	$\lvert\mathcal{X}^*\rvert$	Opus	Grd	Opus	Grd
australian	14 (3/5/6)	690	2	1.00/1.00	5/5	26/25	1/2	0.57/0.56	0.57/0.55
coil2000	85 (0/85/0)	9822	2	0.05/0.05	1/1	1/1	75/38	0.06/0.06	0.14/0.14
fars	29 (5/0/24)	100968	8	0.65/0.65	2/2	4/2	49/43	0.66/0.66	0.68/0.68
german	20 (0/7/13)	1000	2	0.80/1.00	7/6	3040/3065	5/5	0.22/0.21	0.21/0.21
heart	13 (1/12/0)	270	2	1.00/1.00	4/4	8/9	1/2	0.42/0.42	0.43/0.42
ionosphere	33 (32/1/0)	351	2	1.00/1.00	3/3	549/962	1/5	0.61/0.64	0.59/0.66
kddcup	41 (26/0/15)	494020	23	0.95/0.95	2/2	159/122	706/412	0.96/0.97	0.99/0.99
letter	16 (0/16/0)	20000	26	0.95/0.95	5/4	1220/1914	204/122	0.61/0.61	0.62/0.61
lymph.	18 (0/13/5)	148	4	1.00/1.00	4/5	63/85	1/1	0.49/0.48	0.49/0.48
magic	10 (10/0/0)	19020	2	1.00/1.00	5/4	118/435	7/35	0.43/0.43	0.43/0.43
move. libras	90 (90/0/0)	360	15	0.85/0.90	3/2	2043/3183	66/90	0.36/0.36	0.38/0.36
optdigits	64 (0/64/0)	5620	10	0.35/0.45	2/3	16/132	128/122	0.36/0.46	0.59/0.54
pageblocks	10 (4/6/0)	5472	5	1.00/1.00	4/5	58/70	3/8	0.65/0.73	0.65/0.73
penbased	16 (0/16/0)	10992	10	1.00/1.00	5/4	1228/1784	17/28	0.78/0.76	0.78/0.77
ring	20 (20/0/0)	7400	2	0.45/0.35	5/6	1819/777	27/18	0.30/0.35	0.30/0.48
satimage	36 (0/36/0)	6435	7	0.80/0.80	4/4	988/1861	69/104	0.74/0.75	0.73/0.74
segment	19 (19/0/0)	2310	7	1.00/1.00	3/3	153/197	6/9	0.86/0.86	0.86/0.87
sonar	60 (60/0/0)	208	2	0.70/0.70	4/3	435/1808	3/10	0.45/0.41	0.45/0.40
spambase	57 (57/0/0)	4597	2	0.50/0.30	4/3	383/180	15/38	0.50/0.33	0.66/0.57

spectfheart	44 (0/44/0)	267	2	0.65/0.65	3/3	938/1747	4/14	0.34/0.37	0.34/0.33
texture	40 (40/0/0)	5500	11	0.80/0.80	4/4	409/765	46/50	0.76/0.75	0.73/0.77
thyroid	21 (6/15/0)	7200	3	0.55/0.85	3/4	18/24	4/4	0.55/0.85	0.55/0.85
twonorm	20 (20/0/0)	7400	2	0.60/0.20	6/5	1414/200	26/24	0.46/0.20	0.46/0.41
vehicle	18 (0/18/0)	846	4	1.00/1.00	4/3	1275/872	4/10	0.48/0.50	0.49/0.49
vowel	13 (10/3/0)	990	11	1.00/1.00	3/3	12/19	5/5	0.45/0.49	0.47/0.49
wdbc	30 (30/0/0)	569	2	1.00/1.00	5/2	373/286	2/8	0.81/0.82	0.82/0.82
wine	13 (13/0/0)	178	3	1.00/1.00	2/2	1/1	1/1	0.76/0.79	0.74/0.79
wine-red	11 (11/0/0)	1599	11	1.00/1.00	5/4	158/256	12/16	0.21/0.22	0.22/0.23
wine-white	11 (11/0/0)	4898	11	1.00/1.00	5/4	481/779	11/24	0.20/0.19	0.19/0.19
avg.	30 (12/6/2)	25000	7	0.81/0.81	3.8/3.5	599/743	51/43	0.52/0.53	0.53/0.55

6
Conclusion

This dissertation proposed knowledge discovery algorithms to assist data analysts with data exploration, discovering powerful description models, or concluding that no satisfactory models exist, implying therefore new experiments and data are required for the phenomena under investigation. For our solutions to be effective, we put special emphasis in interpretability, statistical robustness, and efficient exact algorithms. This way analysts can understand the results, trust them to represent aspects of the underlying data generating process, and obtain them fast knowing they are indeed the best solutions. Extensive experimental evaluation showed that our estimators have attractive statistical properties, outperforming the state-of-the-art on the intensive discovery tasks considered. The optimization algorithms and bounding functions proposed are very effective in practice, requiring a reasonable amount of time for exact search. The fast greedy algorithms provided near-optimal results on roughly 50 benchmark datasets. Qualitative results from case studies with Materials Science data, Bayesian networks, clustering datasets, and toy examples such as Tic-tac-toe, demonstrated that the products of this dissertation can indeed assist data analysts in discovering knowledge. This statement is supported by Materials Science researchers who corroborated our discoveries. On the same data, the state-of-the-art Markov blanket discovery algorithms and mutual information

estimators were inadequate.

6.1 Discussion and future directions

The general methodology of this dissertation is the following: use information-theoretic measures for model assessment, regularize them to be robust and allow generalization when estimated from observational data, and use combinatorial optimization to find the best model. We follow this methodology in all three problems investigated, namely functional dependency discovery, discovery of totally correlated sets, and partitioning for mixed-data mutual information estimation. There are many advantages with this approach:

- the quality of a model can be assessed and quantified in a meaningful way. That is, information theory gives the tools to construct objective functions tailored to some problem and allows us to argue in terms of (in)dependence, redundancy, and relevancy, in a non-parametric manner. Moreover, we can quantify the quality of a model in intuitive scales.
- The regularization brings robustness which we can model in certain ways to favor specific structures. For example, the permutation model as a regularizer takes into account the data configuration, while the upper-bound (Prop. 3.2.1) only takes into account the marginal counts. Another example is a trade-off between type I and II error, with stricter regularization favoring the former. Moreover, the regularization can turn monotone objective functions into non-monotone, favoring sparsity as well as leading to parameter-free methods.
- Different optimization algorithms can be applied depending on the problem setting. For example, if exact algorithms are not practical, then one can resort to greedy algorithms of various flavors, e.g., standard greedy, stochastic, accelerated, etc., and approximation guarantees can be derived to bound the solution error. In general, advances in algorithms translate to advances in the efficiency and applicability of the knowledge discovery tasks.

That being said, there are possible configurations that we did not consider in this dissertation. For example, we could have investigated different normalizers for the total correlation, e.g., the cardinality of the set, or even no normalization, and then proceed to derive pruning functions. Another example would be to study larger

problem instances and use faster greedy algorithms, e.g., stochastic greedy, and evaluate their performance. Also, we could have derived approximation guarantees for greedy optimization with the regularized scores we considered (see Sec. 3.6.1).

Regarding different knowledge discovery tasks to apply our methodology, one could consider time series analysis. In fact, there exists a direct analogue of mutual information for time series, the directed information [Mar73, Mas90], which possesses good theoretical properties [PKW11] and links to causality [WR19]. One could employ directed information for both supervised [ZS16] and unsupervised discovery tasks, develop efficient discovery algorithms, and propose estimation techniques such as coding distributions and weighting [JPZ+13, BEYY04, WST97, Bel15, Hut13] to be robust against sparsity.

Another possible direction is establishing relations between null-unbiased estimation combined with top-k optimization and significance testing/multiple hypotheses (see Sec. 3.6.2). The statistics and data mining community have contributed a lot towards pattern mining tasks following the latter [TTS13, LLSPB15, Tar90, Ham10, Web07]. Demonstrating the validity of the former could be beneficial in many applications that require scoring and retrieving the top-k, since, for example, one can use approximate algorithms with the possibility of guarantees.

Lastly, one could further investigate connections between score-based and independence-based approaches for Markov blanket discovery (see Sec. 3.6.3). For example, maximizing the fraction of information for top-k minimal requires no assumptions at all and can retrieve multiple Markov blankets. Using an algorithmic framework that is better tailored for large k values, e.g., from Pennerath [Pen18], as well as finding a criterion to determine the number of Markov blankets, could potentially result in a powerful algorithm for multiple MB discovery (see Sec. 3.6.4).

Bibliography

[Ada04] Christoph Adami. Information theory in molecular biology. *Physics of Life Reviews*, 1(1):3–22, April 2004.

[AK01] András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001.

[AST+10] Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D. Koutsoukos. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. *Journal of Machine Learning Research*, 11(7):171–234, 2010.

[ASTB03] C. F. Aliferis, A. R. Statnikov, I. Tsamardinos, and L. E. Brown. Causal Explorer: A Causal Probabilistic Network Learning Toolkit for Biomedical Discovery. In *Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, METMBS*, pages 371–376, December 2003.

[ATS03] C.F. Aliferis, I. Tsamardinos, and A. Statnikov. HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection. *AMIA Annual Symposium Proceedings*, 2003:21–25, 2003.

[Bat94] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, July 1994.

[Bel15] Marc G. Bellemare. Count-based frequency estimation with bounded memory. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3337–3344. AAAI Press, July 2015.

[BEYY04] R. Begleiter, R. El-Yaniv, and G. Yona. On Prediction Using Variable Order Markov Models. *Journal of Artificial Intelligence Research*, 22:385–421, December 2004.

[BMS97] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: generalizing association rules to correlations. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 265–276. Association for Computing Machinery, June 1997.

[BPZL12] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *Journal of Machine Learning Research*, 13(2):27–66, 2012.

[BSCC89] Ingo A. Beinlich, H. J. Suermondt, R. Martin Chavez, and Gregory F. Cooper. The ALARM Monitoring System: A Case Study with two Probabilistic Inference Techniques for Belief Networks. In Jim Hunter, John Cookson, and Jeremy Wyatt, editors, *European Conference on Artificial Intelligence in Medicine*, pages 247–256, Berlin, Heidelberg, 1989. Springer.

[BSY19] Thomas B. Berrett, Richard J. Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *Annals of Statistics*, 47(1):288–318, February 2019.

[Chi03] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, March 2003.

[CL68] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, May 1968.

[CP87] Roger Cavallo and Michael Pittarelli. The Theory of Probabilistic Databases. In *Proceedings of the 13th International Conference on Very Large Data Bases (VLDB), Brighton, UK*, 1987.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[DGP+08] Ding-Zhu Du, Ronald L. Graham, Panos M. Pardalos, Peng-Jun Wan, Weili Wu, and Wenbo Zhao. Analysis of greedy approximations with non-submodular potential functions. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 167–175, San Francisco, California, January 2008. Society for Industrial and Applied Mathematics.

[DV99] G.A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, May 1999.

[FH61] Robert M. Fano and David Hawkins. Transmission of Information: A Statistical Theory of Communications. *American Journal of Physics*, 29(11):793–794, November 1961. Publisher: American Association of Physics Teachers.

[FI93] U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conferences on Artificial Intelligence*, 1993.

[Fle04] François Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5:1531–1555, 2004.

[FMV07] Uriel Feige, Vahab S. Mirrokni, and Jan Vondrak. Maximizing Non-Monotone Submodular Functions. In *48th Annual IEEE Symposium on Foundations of Computer Science*, pages 461–471, October 2007. ISSN: 0272-5428.

[FS18] Pasi Fränti and Sami Sieranoja. K-means properties on six clustering benchmark datasets. *Applied Intelligence*, 48(12):4743–4759, December 2018.

[GAE07] Isabelle Guyon, Constantin Aliferis, and André Elissee. Causal Feature Selection. In Hiroshi Motoda and Huan Liu, editors, *Computational*

Methods of Feature Selection, volume 20071386, pages 63–85. Chapman and Hall/CRC, October 2007. Series Title: Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

[GBV+17] Bryan R. Goldsmith, Mario Boley, Jilles Vreeken, Matthias Scheffler, and Luca M. Ghiringhelli. Uncovering structure-property relationships of materials by subgroup discovery. *New Journal of Physics*, 19(1):013031, January 2017. Publisher: IOP Publishing.

[GE03] Isabelle Guyon and André Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3(Mar):1157–1182, 2003.

[GH94] Dan Geiger and David Heckerman. Learning Gaussian networks. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 235–243, San Francisco, CA, USA, July 1994. Morgan Kaufmann Publishers Inc.

[GKOV17] Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Estimating mutual information for discrete-continuous mixtures. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5988–5999, Long Beach, California, USA, December 2017. Curran Associates Inc.

[GR04] Chris Giannella and Edward Robertson. On approximation measures for functional dependencies. *Information Systems*, 29(6):483–507, September 2004.

[Gra88] Peter Grassberger. Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, 128(6):369–373, April 1988.

[Gra08] P. Grassberger. Entropy Estimates from Insufficient Samplings. *arXiv:physics/0307138*, January 2008. arXiv: physics/0307138.

[GS11] Jelle J. Goeman and Aldo Solari. Multiple Testing for Exploratory Research. *Statistical Science*, 26(4):584–597, November 2011. Publisher: Institute of Mathematical Statistics.

[GSG15] Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Estimating mutual information by local Gaussian approximation. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 278–287, Amsterdam, Netherlands, July 2015. AUAI Press.

[GVL+15] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler. Big Data of Materials Science: Critical Role of the Descriptor. *Physical Review Letters*, 114(10):105503, March 2015. Publisher: American Physical Society.

[Ham10] Wilhelmiina Hamalainen. Efficient Discovery of the Top-K Optimal Dependency Rules with Fisher's Exact Test of Significance. In *2010 IEEE International Conference on Data Mining*, pages 196–205, December 2010.

[HFEM07] H. Heikinheimo, M. Fortelius, J. Eronen, and H. Mannila. Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters. *Journal of Biogeography*, 34(6):1053–1064, 2007.

[HJM+09] Patrik O. Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 689–696. Curran Associates, Inc., 2009.

[HMC06] David Heckerman, Christopher Meek, and Gregory Cooper. A Bayesian Approach to Causal Discovery. In Dawn E. Holmes and Lakhmi C. Jain, editors, *Innovations in Machine Learning: Theory and Applications*, pages 1–28. Springer, Berlin, Heidelberg, 2006.

[HR70] M. Hellman and J. Raviv. Probability of error, equivocation, and the Chernoff bound. *IEEE Transactions on Information Theory*, 16(4):368–372, July 1970.

[HS16] Thibaut Horel and Yaron Singer. Maximization of approximately submodular functions. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3053–3061, Barcelona, Spain, December 2016. Curran Associates Inc.

[HTT09] Tony Hey, Stewart Tansley, and Kristin Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, October 2009. Publication Title: The Fourth Paradigm: Data-Intensive Scientific Discovery.

[Hut13] Marcus Hutter. Sparse Adaptive Dirichlet-Multinomial-like Processes. In *Conference on Learning Theory*, pages 432–459. PMLR, June 2013.

[HW19] Wilhelmiina Hämäläinen and Geoffrey I. Webb. A tutorial on statistically sound pattern discovery. *Data Mining and Knowledge Discovery*, 33(2), 2019.

[JKP94] George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant Features and the Subset Selection Problem. In *Machine Learning Proceedings 1994*, pages 121–129. Elsevier, 1994.

[JPZ+13] Jiantao Jiao, Haim H. Permuter, Lei Zhao, Young-Han Kim, and Tsachy Weissman. Universal Estimation of Directed Information. *IEEE Transactions on Information Theory*, 59(10):6220–6242, October 2013.

[JVHW15] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax Estimation of Functionals of Discrete Distributions. *IEEE Transaction on Information Theory*, 61(5):2835–2885, May 2015.

[KA14] Justin B. Kinney and Gurinder S. Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences of the United States of America*, 111(9):3354–3359, March 2014.

[KCN08] Yiping Ke, James Cheng, and Wilfred Ng. Correlated pattern mining in quantitative databases. *ACM Transactions on Database Systems*, 2008.

[KG05] Andreas Krause and Carlos Guestrin. Near-optimal Nonmyopic Value of Information in Graphical Models. *Proceedings of the 21st International Conference on Uncertainty in Artificial Intelligence, Cambridge, MA*, 2005.

[KJ97] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, December 1997.

[KL51] S. Kullback and R. A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, March 1951. Publisher: Institute of Mathematical Statistics.

[KLLZ16] Henning Köhler, Uwe Leck, Sebastian Link, and Xiaofang Zhou. Possible and certain keys for SQL. *The VLDB Journal*, 25(4):571–596, August 2016.

[KPHN16] Sebastian Kruse, Thorsten Papenbrock, Hazar Harmouch, and Felix Naumann. Data anamnesis: Admitting raw data into an organization. *IEEE Data Engineering Bulletin*, 39:8–20, 06 2016.

[KR92] Kenji Kira and Larry A. Rendell. The feature selection problem: traditional methods and a new algorithm. In *Proceedings of the tenth national conference on Artificial intelligence*, pages 129–134, San Jose, California, July 1992. AAAI Press.

[KSG04] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, June 2004. Publisher: American Physical Society.

[KV12] Bernhard Korte and Jens Vygen. *Combinatorial Optimization: Theory and Algorithms*. Springer Publishing Company, Incorporated, 5th edition, 2012.

[Lan69] H. O Lancaster. *The chi-squared distribution*. Wiley, New York, 1969. OCLC: 29301.

[Lew92] David D. Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of the workshop on Speech and Natural Language*, pages 212–217, USA, February 1992. Association for Computational Linguistics.

[Lin88] R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, March 1988. Conference Name: Computer.

[LLLC12] Jixue Liu, Jiuyong Li, Chengfei Liu, and Yongfeng Chen. Discover Dependencies from Data—A Review. *IEEE Transactions on Knowledge and Data Engineering*, 24(2):251–264, February 2012. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[LLSPB15] Felipe Llinares-López, Mahito Sugiyama, Laetitia Papaxanthos, and Karsten Borgwardt. Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 725–734, New York, NY, USA, August 2015. Association for Computing Machinery.

[LLZ10] Huawen Liu, Lei Liu, and Huijie Zhang. Ensemble gene selection by grouping for microarray data classification. *Journal of Biomedical Informatics*, 43(1):81–87, February 2010.

[LMSZ10] Huan Liu, Hiroshi Motoda, Rudy Setiono, and Zheng Zhao. Feature selection: An ever evolving frontier in data mining. In Huan Liu, Hiroshi Motoda, Rudy Setiono, and Zheng Zhao, editors, *Proceedings of Machine Learning Research*, volume 10, pages 4–13, 2010.

[LT06] Dahua Lin and Xiaoou Tang. Conditional infomax learning: An integrated framework for feature extraction and fusion. In *European Conference on Computer Vision*, pages 68–82. Springer-Verlag, 2006.

[LY05] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502, April 2005.

[Mar73] H. Marko. The Bidirectional Communication Theory - A Generalization of Information Theory. *IEEE Transactions on Communications*, 21(12):1345–1351, December 1973.

[Mas90] J. Massey. Causality, feedback and directed information. *Proceedings of the 1990 International Symposium on Information Theory and its Applications*, pages 27–30, 1990.

[MBK+15] Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrák, and Andreas Krause. Lazier than lazy greedy. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1812–1818, Austin, Texas, January 2015. AAAI Press.

[MBSP03] Obdulia Molina, Richard P. Brown, Nicolás M. Suárez, and José J. Pestano. The origin of the Osorian shrew (Crocidura osorio) from Gran Canaria resolved using mtDNA. *Italian Journal of Zoology*, 70(2):179–181, January 2003.

[MBV17] Panagiotis Mandros, Mario Boley, and Jilles Vreeken. Discovering Reliable Approximate Functional Dependencies. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 355–363, Halifax, NS, Canada, August 2017. Association for Computing Machinery.

[MBV18] Panagiotis Mandros, Mario Boley, and Jilles Vreeken. Discovering Reliable Dependencies from Data: Hardness and Improved Algorithms. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 317–326, November 2018. ISSN: 2374-8486.

[MBV19] Panagiotis Mandros, Mario Boley, and Jilles Vreeken. Discovering Reliable Correlations in Categorical Data. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1252–1257, November 2019. ISSN: 2374-8486.

[MBV20] Panagiotis Mandros, Mario Boley, and Jilles Vreeken. Discovering dependencies with reliable mutual information. *Knowledge and Information Systems*, pages 1–31, 2020.

[MCGBK19] Francisco J. Montáns, Francisco Chinesta, Rafael Gómez-Bombarelli, and J. Nathan Kutz. Data-driven modeling and learning in science and engineering. *Comptes Rendus Mécanique*, 347(11):845–855, November 2019.

[MHH+01] Renée J. Miller, Mauricio A. Hernández, Laura M. Haas, Lingling Yan, C. T. Howard Ho, Ronald Fagin, and Lucian Popa. The clio project: Managing heterogeneity. *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, 30(1):78–83, March 2001.

[Min78] Michel Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In J. Stoer, editor, *Optimization Techniques*, pages 234–243, Berlin, Heidelberg, 1978. Springer.

[MKBV20] Panagiotis Mandros, David Kaltenpoth, Mario Boley, and Jilles Vreeken. Discovering functional dependencies from mixed-type data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1404–1414, 2020.

[MR89] Christopher J. Matheus and Larry A. Rendell. Constructive induction on decision trees. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 645–650. Morgan Kaufmann, 1989.

[MS08] Kurt Mehlhorn and Peter Sanders. *Algorithms and Data Structures: The Basic Toolbox*. Springer-Verlag, Berlin Heidelberg, 2008.

[MT99] Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pages 505–511, Cambridge, MA, USA, November 1999. MIT Press.

[NEB09] Vinh Nguyen, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the Twenty-Sixth International Conference on Machine Learning*, pages 1073–1080. Association for Computing Machinery (ACM), 2009.

[NEB10] Xuan Vinh Nguyen, Julien Epps, and James Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 2010.

[NK04] Ullas Nambiar and Subbarao Kambhampati. Mining approximate functional dependencies and concept similarities to answer imprecise queries. In *Proceedings of the 7th International Workshop on the Web and Databases: Colocated with ACM SIGMOD/PODS 2004*, page 73–78, New York, NY, USA, 2004. Association for Computing Machinery.

[NMV16] Hoang-Vu Nguyen, Panagiotis Mandros, and Jilles Vreeken. Universal Dependency Analysis. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 792–800. Society for Industrial and Applied Mathematics, June 2016.

[NSB02] Ilya Nemenman, F. Shafee, and William Bialek. Entropy and Inference, Revisited. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 471–478. MIT Press, 2002.

[Omi03] E.R. Omiecinski. Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57–69, January 2003.

[Pan03] Liam Paninski. Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253, June 2003. Publisher: MIT Press.

[Pan04] L. Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50(9):2200–2203, September 2004.

[PE08] Jean-Philippe Pellet and André Elisseeff. Using Markov Blankets for Causal Structure Learning. *Journal of Machine Learning Research*, 9(Jul):1295–1342, 2008.

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[Pea09] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, 2009.

[PEM+15] Thorsten Papenbrock, Jens Ehrlich, Jannik Marten, Tommy Neubert, Jan-Peer Rudolph, Martin Schönberg, Jakob Zwiener, and Felix Naumann. Functional dependency discovery: An experimental evaluation of seven algorithms. *Proceedings of the VLDB Endowment*, 8(10):1082–1093, June 2015.

[Pen10] Frédéric Pennerath. Fast extraction of locally optimal patterns based on consistent pattern function variations. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, volume 6323, pages 34–49. Springer, 2010.

[Pen18] Frédéric Pennerath. An efficient algorithm for computing entropic measures of feature subsets. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 483–499. Springer, 2018.

[PKW11] Haim H. Permuter, Young-Han Kim, and Tsachy Weissman. Interpretations of Directed Information in Portfolio Theory, Data Compression, and Hypothesis Testing. *IEEE Transactions on Information Theory*, 57(6):3248–3259, June 2011.

[PLD05] Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, August 2005. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[PMJS14] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, January 2014.

[PN17] Thorsten Papenbrock and Felix Naumann. Data-driven schema normalization. In *20th International Conference on Extending Database Technology*, 2017.

[PNBT07] Jose M. Peña, Roland Nilsson, Johan Björkegren, and Jesper Tegnér. Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning*, 45(2):211–232, July 2007.

[PY08] Liam Paninski and Masanao Yajima. Undersmoothed Kernel Entropy Estimators. *IEEE Transactions on Information Theory*, 54(9):4384–4388, September 2008.

[Qui86] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, March 1986.

[RBNV14] Simone Romano, James Bailey, Vinh Nguyen, and Karin Verspoor. Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance. In *International Conference on Machine Learning*, pages 1143–1151, January 2014. ISSN: 1938-7228 Section: Machine Learning.

[RG99] Raghu Ramakrishnan and Johannes Gehrke. *Database Management Systems*. McGraw-Hill, Inc., USA, 2nd edition, 1999.

[Rou99] Mark S Roulston. Estimating the errors on measured entropy and mutual information. *Physica D: Nonlinear Phenomena*, 125(3):285–294, January 1999.

[RRF+11] David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. Detecting Novel Associations in Large Data Sets. *Science*, 334(6062):1518–1524, December 2011.

[RVBV16] Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. A Framework to Adjust Dependency Measure Estimates for Chance. In

Proceedings of the 2016 SIAM International Conference on Data Mining, pages 423–431. Society for Industrial and Applied Mathematics, June 2016.

[SB13] Mahito Sugiyama and Karsten M. Borgwardt. Measuring statistical dependence via the mutual information dimension. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1692–1698, Beijing, China, August 2013. AAAI Press.

[SBG+20] Christopher Sutton, Mario Boley, Luca M. Ghiringhelli, Matthias Rupp, Jilles Vreeken, and Matthias Scheffler. Identifying domains of applicability of machine learning models for materials science. *Nature Communications*, 11(1):4428, September 2020.

[Sch13] Steffen Schober. Some worst-case bounds for Bayesian estimators of discrete distributions. In *2013 IEEE International Symposium on Information Theory*, pages 2194–2198, July 2013. ISSN: 2157-8117.

[SG96] Thomas Schürmann and Peter Grassberger. Entropy estimation of symbol sequences. *Chaos*, 6(3):414–427, September 1996. Publisher: American Institute of Physics.

[SG14] Greg Ver Steeg and Aram Galstyan. Discovering structure in high-dimensional data through Correlation Explanation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, pages 577–585, Montreal, Canada, December 2014. MIT Press.

[SGS93] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.

[Sha48] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[SLA13] Alexander Statnikov, Jan Lemeir, and Constantin F. Aliferis. Algorithms for discovery of multiple Markov boundaries. *Journal of Machine Learning Research*, 14(1):499–566, February 2013.

[SN10] Jorge Silva and Shrikanth Narayanan. Nonproduct Data-Dependent Partitions for Mutual Information Estimation: Strong Consistency and Applications. *IEEE Transactions on Signal Processing*, 58(7):3497–3511, July 2010.

[SRAFFH+11] Luciano Sánchez Ramos, Jesús Alcalá Fernández, Alberto Fernández Hilario, Julián Luengo Martín, Joaquín Derrac Rus, Salvador García López, and Francisco Herrera Triguero. KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 2011.

[Suz16] Joe Suzuki. An Estimator of Mutual Information and its Application to Independence Testing. *Entropy*, 18(4):109, April 2016.

[Suz19] Joe Suzuki. Mutual Information Estimation: Independence Detection and Consistency. In *2019 IEEE International Symposium on Information Theory*, pages 2514–2518, July 2019. ISSN: 2157-8117.

[TA03] Ioannis Tsamardinos and Constantin Aliferis. Towards Principled Feature Selection: Relevancy, Filters and Wrappers. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Morgan Kaufmann Publishers, 2003.

[Tar90] R. E. Tarone. A Modified Bonferroni Method for Discrete Data. *Biometrics*, 46(2):515–522, 1990. Publisher: [Wiley, International Biometric Society].

[TAS03] Ioannis Tsamardinos, Constantin F. Aliferis, and Alexander Statnikov. Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 673–678, 2003.

[TASS03] Ioannis Tsamardinos, Constantin Aliferis, Alexander Statnikov, and Er Statnikov. Algorithms for Large Scale Markov Blanket Discovery. In *In The 16th International FLAIRS Conference, St*, pages 376–380. AAAI Press, 2003.

[TD68] D. Tebbe and S. Dwyer. Uncertainty and the probability of error (Corresp.). *IEEE Transactions on Information Theory*, 14(3):516–518, May 1968.

[Tib96] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[TL18] Nicholas M. Timme and Christopher Lapish. A Tutorial for Information Theory in Neuroscience. *eNeuro*, 5(3):ENEURO.0052–18.2018, May 2018.

[TPB00] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[TTS13] Aika Terada, Koji Tsuda, and Jun Sese. Fast Westfall-Young permutation procedure for combinatorial regulation discovery. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pages 153–158, December 2013.

[Tuk77] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[Tuk80] John W. Tukey. We Need Both Exploratory and Confirmatory. *The American Statistician*, 34(1):23–25, 1980.

[VCB14] Nguyen Xuan Vinh, Jeffrey Chan, and James Bailey. Reconsidering Mutual Information Based Feature Selection: A Statistical Significance View. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, June 2014.

[VE14] Jorge R. Vergara and Pablo A. Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1):175–186, January 2014.

[VV69] J. A. Van Vechten. Quantum Dielectric Theory of Electronegativity in Covalent Systems. I. Electronic Dielectric Constant. *Physical Review*, 182(3):891–905, June 1969. Publisher: American Physical Society.

[VV11] Gregory Valiant and Paul Valiant. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 685–694, San Jose, California, USA, June 2011. Association for Computing Machinery.

[VV13] Paul Valiant and Gregory Valiant. Estimating the Unseen: Improved Estimators for Entropy and other Properties. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2157–2165. Curran Associates, Inc., 2013.

[VWI97] Paul Viola and William M. Wells III. Alignment by Maximization of Mutual Information. *International Journal of Computer Vision*, 24(2):137–154, September 1997.

[Wat60] Satosi Watanabe. Information Theoretical Analysis of Multivariate Correlation. *IBM Journal of Research and Development*, 4(1):66–82, January 1960.

[Web95] G. I. Webb. OPUS: An Efficient Admissible Algorithm for Unordered Search. *Journal of Artificial Intelligence Research*, 3:431–465, December 1995.

[Web07] Geoffrey I. Webb. Discovering Significant Patterns. *Machine Learning*, 68(1):1–33, July 2007.

[Web08] Geoffrey I. Webb. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. *Machine Learning*, 71(2-3):307–323, June 2008.

[WF03] Abraham Wyner and Dean Foster. On the lower limits of entropy estimation. *IEEE Transactions on Information Theory*, 07 2003.

[WR19] Aleksander Wieczorek and Volker Roth. Information Theoretic Causal Effect Quantification. *Entropy*, 21(10):975, October 2019.

[WRN+17] Yisen Wang, Simone Romano, Vinh Nguyen, James Bailey, Xingjun Ma, and Shu-Tao Xia. Unbiased Multivariate Correlation Analysis. In *Thirty-First AAAI Conference on Artificial Intelligence*, February 2017.

[WST97] Frans Willems, Yuri Shtarkov, and Tjalling Tjalkens. Tj.J.: Reflections on 'The Context-Tree Weighting Method: Basic Properties. *Newsletter of the IEEE Information Theory Society*, 1997.

[WY16] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *arXiv:1407.0381 [cs, math, stat]*, February 2016. arXiv: 1407.0381.

[XTK06] Hui Xiong, Pang-Ning Tan, and Vipin Kumar. Hyperclique pattern discovery. *Data Mining and Knowledge Discovery*, 2006.

[YM99] Howard Hua Yang and John Moody. Feature selection based on joint mutual information. In *Proceedings of the 12th Annual Conference on Neural Information Processing Systems*, pages 22–25, 1999.

[ZPWN08] Xiang Zhang, Feng Pan, Wei Wang, and Andrew Nobel. Mining Non-Redundant High Order Correlations in Binary Data. *Proceedings of the VLDB Endowment*, 1(1):1178–1188, August 2008.

[ZS16] Yuxun Zhou and Costas J Spanos. Causal meets Submodular: Subset Selection with Directed Information. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2649–2657. Curran Associates, Inc., 2016.

Glossary

b_0 . 29
m_0 . 28
$m_{\bar{0}}$. 77
$m_{\bar{\bar{0}}}$. 77
$t_{\bar{0}}$. 77
$t_{\bar{\bar{0}}}$. 78

admissible bounding function . 40
alphabetical refinement operator . 45

base transformation . 35
Bayesian network . 19
branch-and-bound . 45
branch-informed bounding function . 83
branch-informed upper-bound . 83

chain bounding function . 44, 84
chance-adjusted estimator . 51
chi-square estimator . 51
conditional mutual information . 11
conditional Shannon entropy . 10
consistent . 14
contrained optimal partition . 107
converging strategies . 100

decreasing-entropy refinement operator . 80
diminishing returns property . 47
discretization strategy . 100
dominated convergence . 100

empirical counts . 13
empirical distribution . 13
extended transformation . 38

faithfulness . 20
fraction of information . 11
functional dependency . 11
functional dependency discovery . 23

greedy algorithm . 46

homomorphic relation . 33

labeling . 13
low entropy extension . 80

Markov blanket . 20
MDL estimator . 51
MIC estimator . 110
minimax estimator . 14
monotonicity bounding function . 40, 82

mutual information . 10

normalized total correlation . 75

optimal substructure . 107
optimistic estimator . 40
OPUS . 46

permutation fraction of information . 28
permutation model . 27
permutation mutual information . 28
permutation normalized total correlation . 79
plugin . 14
positive gain . 41
potential . 40

quantization strategies . 99

refinement operator . 45
regret . 85

sample complexity . 14
Shannon entropy . 9
shrink step . 49
specialization bounding function . 41
specialization relation . 31
statistical independence . 10
submodular . 47

total correlation . 74
total correlation upper-bound . 75

unseen estimator . 14