entropy

MDPI

Article

Information Theoretic Methods for Variable Selection—A Review

Jan Mielniczuk [1,2]

[1] Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warsaw, Poland; j.mielniczuk@ipipan.waw.pl
[2] Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland

Abstract: We review the principal information theoretic tools and their use for feature selection, with the main emphasis on classification problems with discrete features. Since it is known that empirical versions of conditional mutual information perform poorly for high-dimensional problems, we focus on various ways of constructing its counterparts and the properties and limitations of such methods. We present a unified way of constructing such measures based on truncation, or truncation and weighing, for the Möbius expansion of conditional mutual information. We also discuss the main approaches to feature selection which apply the introduced measures of conditional dependence, together with the ways of assessing the quality of the obtained vector of predictors. This involves discussion of recent results on asymptotic distributions of empirical counterparts of criteria, as well as advances in resampling.

Keywords: conditional independence; interaction information; Möbius expansion; Markov blanket; feature selection

Citation: Mielniczuk, J. Information Theoretic Methods for Variable Selection—A Review. *Entropy* **2022**, *24*, 1079. https://doi.org/10.3390/e24081079

Academic Editor: Ciprian Doru Giurcaneanu

Received: 24 June 2022
Accepted: 2 August 2022
Published: 4 August 2022

1. Introduction

Conditional independence is one of the main concepts in statistics which plays a fundamental role in such areas as causal inference, dependence analysis, and graphical modelling. In this review, we discuss how measures of dependence considered in information theory are used in classification and regression problems to choose predictors which significantly influence the outcome. This is a vital application of the information theoretic approach to variable selection, also known as feature selection. Let us stress, however, that information-theoretic measures are frequently applied in many other problems in Machine Learning and Statistics, in contexts other than feature selection. Some representative examples include data visualisation (t-SNE, [1]), clustering [2], Independent Component Analysis (ICA [3], Chapter 15), Variational Inference ([4], Chapter 10), and Natural Language Modelling [5], among others.

In the times of the big data challenge, the problem of a choice of a small group of variables from the pool of all potential variables, all of which would accurately describe the changes in response, is gaining in importance. This is needed for better understanding of a studied phenomena, as well as for construction of tractable models for them. Moreover, feature selection is instrumental in avoiding curse of dimensionality problem when a prohibitive amount of data are required for adequate fitting of the model ([6], Chapter 2) and avoiding overfitting (ibid., Chapter 7). This is also important for prediction, as classifiers which avoid using inactive predictors are usually less variable. Feature selection is frequently applied because of cost considerations, especially when measuring some of the potentially useful predictors is costly or inconvenient.

Another competing direction of research with a similar aim is dimensionality reduction, in particular variable extraction, which transforms given variables to obtain a lean group of new predictors. Primary examples of such methods are Principal Components Regression

and Partial Least Squares, see, e.g., [6], and more recent approaches based on neural networks [7].

Here, we focus on an important group of selectors, called *filters*, which, during the selection process, do not take into account the classifier (or regression estimator) which will incorporate selected variables. In contrast, *wrappers* select features to optimise performance of a specific classifier under consideration (see, e.g., methods which also yield ranking of features, such as [8–10]). The property that filters are model-agnostic makes them a universal tool which can be used for any classifier or regression estimator.

In this paper, we try to present critical and, unavoidably, partial assessment of the state of the art for this problem, focusing on results which can be formally proved, and showing motivation, advantages, and limitations of the presented solutions, including the most recent ones. Additionally, we show that the majority of information-based feature selection criteria can be viewed as truncated or truncated and weighted expansions of Möbius decomposition. This sheds new light on similarities and differences between the frequently employed criteria. Moreover, based on the recent results, asymptotic distributions of their plug-in empirical counterparts are discussed. Such results are needed to construct a test of conditional independence which control probability of false alarms. The recent advances in resampling yielding conditionally independent samples which provide alternative method to solve this problem are also discussed. By this, hopefully, a new insight is added to existing reviews (see e.g., [11–14]).

The paper is organised as follows. We first discuss in Section 2 the main information theoretic objects and their interplay, focusing on Möbius decomposition of conditional mutual information (*CMI*). In Section 3, we introduce and discuss a feature selection problem from the information theory perspective based on measures of dependence introduced in the previous section. The concept of Markov Blanket of a target variable, being the aim of feature selection as the minimal set of predictors containing the whole information about it, is investigated here. In Section 4, we study feature selection criteria related to *CMI* stressing that most of them are naturally related to Möbius decomposition. Moreover, this section introduces another group of selectors based on variational bounds of information measures. In the following Section 5, we discuss the interplay between various *CMI*-related measures. The next sections cover the approximate distributions of the introduced measures (Section 6) and an alternative method of assessing their distributions using resampling schemes (Section 7). These properties are vital for performance of conditional independence tests, which are building blocks of feature selection algorithms. In Section 8, Markov Blanket discovery algorithms are described. Section 9 discusses feature selection in continuous case, Section 10 covers related problem of interaction detection stressing how these can be incorporated into feature selection approach.

2. Conditional Mutual Information and Related Measures

We start with discussing properties and the role that conditional mutual information (*CMI*) plays in variable selection based on information theoretic concepts. We will focus on a discrete finite case, meaning that the considered variables take a finite number of discrete values. The continuous case is reviewed shortly in Section 9. In the following, Y will denote the class variable whereas X and Z (possibly multivariate) are features which will be used to predict Y. For basic information theoretic concepts we refer to [15,16].

2.1. Mutual Information MI

Definition 1 ([15], p. 46)**.** *The Mutual Information (MI) between Y and X is defined as*

$$I(Y;X) = \sum_{x,y} P(X=x, Y=y) \log \frac{P(X=x, Y=y)}{P(X=x)P(Y=y)} = H(Y) - H(Y|X), \tag{1}$$

where $H(Y) = -\sum_y P(Y=y) \log P(Y=y)$ and $H(Y|X) = \sum_x P(X=x) H(Y|X=x)$ are the entropy and the conditional entropy, respectively.

The second equality in (1) motivates the name *Information Gain* also used for MI. It also yields intuitive meaning for MI as the decrease in variability of Y is measured by its entropy when the information about another variable X is available. We remark that $I(Y;X)$ is frequently denoted as $MI(Y;X)$. Note that the semicolon in $I(Y;X)$ determines between which variables dependence is considered: either between Y and X in the case of $I(Y;X)$ or between Y and (X,Z) in the case of $I(Y;X,Z)$. $I(Y;X)$ evaluates how similar the joint distribution P_{YX} of (Y,X) is to the product $P_Y \otimes P_X$ of their marginal distributions. As $P_Y \otimes P_X$ corresponds to independence of X and Y, $I(Y;X)$ can be considered a measure of strength of dependence between Y and X. If follows from the definition that ([15], Section 2.3)

$$I(Y;X) = KL(P_{Y,X}||P_Y \otimes P_X),$$

where *Kullback–Leibler (KL) divergence* between distributions P_W and P_Z is defined as

$$KL(P_W||P_Z) = \sum_w P(W=w)\log\{P(W=w)/P(Z=w)\}.$$

We note that KL divergence is closely related to the Maximum Likelihood (ML) method and popular feature selection method Akaike Information Criterion (AIC) is derived as the bias-corrected ML [17]. We also remark that Kullback–Leibler divergence can be replaced by other pseudo-distance between probability distributions resulting in a different measure of dependence.

It follows from the properties of KL divergence that $I(Y;X)$ is non-negative and is equal to zero if, and only if, $P_{Y,X} = P_Y \otimes P_X$, i.e., when Y and X are independent. Moreover, the definition (1) that Mutual Information is symmetric: $I(Y;X) = I(X;Y)$. It is also easily seen that (see formula (2.450) in [15])

$$I(Y;X) = H(Y) + H(X) - H(Y,X). \tag{2}$$

2.2. Conditional Mutual Information CMI

We denote by $P_{X|Z=z}$ conditional distribution of X given $Z=z$ and define $I(Y;X|Z=z) = KL(P_{Y,X|Z=z}||P_{Y|Z=z} \otimes P_{X|Z=z})$ as the strength of dependence between Y and X given $Z=z$. Thus, $I(Y;X|Z=z)=0$ means that Y and X are independent given that Z equals z. Now we define conditional mutual information.

Definition 2 ([15], p. 49)**.** *The conditional mutual information (CMI) is*

$$\begin{aligned} I(Y;X|Z) &= E_{Z=z} I(Y;X|Z=z) \\ &= \sum_z P(Z=z) \sum_{x,y} P(X=x,Y=y|Z=z) \log \frac{P(X=x,Y=y|Z=z)}{P(X=x|Z=z)P(Y=y|Z=z)}. \end{aligned} \tag{3}$$

Thus, the conditional mutual information is the mutual information of Y and X given $Z=z$ averaged over the values of Z. It is measure of dependence between Y and X given the knowledge of X. From the properties of Kullback–Leibler divergence it follows that

$$I(Y;X|Z) = 0 \iff X \text{ and } Y \text{ are conditionally independent given } Z.$$

This is a powerful property, not satisfied for other measures for dependence, such as partial correlation coefficient in the case of continuous random variables. The conditional independence of X and Y given Z will be denoted by $X \perp\!\!\!\perp Y|Z$ and abbreviated to CI. We note that since $I(Y;X|Z)$ is defined as a probabilistic average of $I(Y;X|Z=z)$ over $Z=z$, it follows that

$$I(Y;X|Z) = 0 \iff I(Y;X|Z=z) = 0 \text{ for any } z \text{ in the support of } Z.$$

Thus, formally, testing conditional independence of X and Y given Z is equivalent to testing unconditional dependence of X and Y on every stratum $Z = z$. This, however, does not make the problem an easy task, as sometimes claimed, since performing such tests simultaneously on each strata (global null), even when we have sufficient data to do it, would result in lack of control of the probability of false signals (multiple testing problem).

The above definitions can be naturally extended to the case of random vectors (i.e., when X, Y, and Z are multivariate) by using a multivariate probability mass functions instead of a univariate one. It is also easily seen that the following chain formula holds

$$I(Y; X, Z) = I(Y; X) + I(Y; Z|X). \tag{4}$$

2.3. Interaction Information II

The 3-way interaction information $II(Y; X, Z)$ of, possibly multivariate, X, Y, and Z plays also important role in feature selection.

Definition 3. *The 3-way interaction information $II(Y; X, Z)$ is defined as*

$$II(Y; X; Z) = I(Y; X|Z) - I(Y; X) = I(Y; X, Z) - I(Y; X) - I(Y; Z). \tag{5}$$

The second equality, stemming from (4), neatly explains the idea of interaction information: we want to evaluate the synergistic effect (positive or negative) of both X and Z influencing Y, disregarding the sum of their individual effects. Thus, from the evaluated strength of *overall* dependence between Y and a pair (X, Z) measured by $I(Y; X, Z)$, we subtract the strength of *individual* dependences of Y on X and Y on Z measured by $I(Y; X)$ and $I(Y; Z)$, respectively. Although it is not immediately clear from the definition, $II(Y; X; Z)$ is actually a symmetric function of its arguments. However, it can be positive, negative, or 0. If it is positive, X and Z are said to be complementary wrt to Y or interacting positively in influencing Y. A smaller than 0 value of II indicates redundancy among features or their inhibition.

We define 2-way as II as

$$II(Y; X) = I(Y; X) \quad \text{and} \quad II(Y; X|Z) = I(Y; X|Z).$$

Definition (5) can be generalised in a recursive way to define k-way interaction information (see [18–20] for alternative equivalent definition). For any subset of indices $S = \{s_1, s_2 \ldots, s_{|S|}\}$, $Z_S := (Z_{s_1}, Z_{s_2}, \ldots, Z_{s_{|S|}})$ will stand for the sub-vector of Z_is with indices in S. Abusing the notion, $Z \in Z_S$ will be mean that $Z \in \{Z_{s_1}, Z_{s_2} \ldots, Z_{s_{|S|}}\}$. Let $S = \{1, \ldots, |S|\}$ and $k = |S|$ be the cardinality of S. We define k-way interaction information $II(Z_1; Z_2; \ldots; Z_{|S|})$ so that the chain formula holds.

Definition 4. *k-way interaction information $II(Z_1; Z_2; \ldots; Z_{|S|})$ is defined as*

$$II(Z_1; Z_2; \ldots; Z_{|S|}) = II(Z_1; \ldots; Z_{|S|-1}|Z_{|S|}) - II(Z_1; \ldots; Z_{|S|-1}), \tag{6}$$

where, consistently with the definition of the conditional mutual information in (3), we define

$$II(Z_1; \ldots; Z_{|S|-1}|Z_{|S|}) = \sum_{z_{|S|}} p(z_{|S|}) II(Z_1; \ldots; Z_{|S|-1}|Z_{|S|} = z_{|S|}).$$

Using expansions of MI and CMI in terms of entropies, the following formula for k-way II is obtained

$$II(Z_1; \ldots; Z_{|S|}) = -\sum_{i=1}^{|S|} \sum_{T \subseteq S, |T|=i} (-1)^{|S|-|T|} H(Z_T), \tag{7}$$

where $Z_T = (Z_{t_1}, \ldots, Z_{t_i})$ for $T = \{t_1, \ldots, t_i\}$. In particular, we have

$$II(Z_1; Z_2; Z_3) = H(Z_1, Z_2) + H(Z_2, Z_3) + H(Z_1, Z_3) - H(Z_1, Z_2, Z_3) - H(Z_1) - H(Z_2) - H(Z_3).$$

Note that when, e.g., $Z_3 = XOR(Z_1, Z_2) = I\{Z_1 \neq Z_2\}$ and (Z_1, Z_2) are independent copies of a random variable having Bernoulli distribution with $p = 1/2$, we obtain $II(Z_1; Z_2; Z_3) = \log 2$, as in this case $I(Z_3, Z_2|Z_1) = H(Z_3)$. In the following, so-called Möbius expansion (see, e.g., [20]) plays a crucial role.

Theorem 1. *The conditional mutual information satisfies the equation*

$$I(X; Y|Z_S) = I(Y; X|Z_1, \ldots, Z_{|S|}) = \sum_{i=0}^{|S|} \sum_{T \subseteq S, |T|=i} II(X; Y; Z_{t_1}, \ldots; Z_{t_i}), \tag{8}$$

where $T = \{t_1, \ldots, t_i\}$ *and, conversely,*

$$II(X; Y; Z_{s_1}, \ldots, Z_{s_{|S|}}) = \sum_{i=0}^{|S|} \sum_{T \subseteq S, |T|=i} (-1)^{|S|-|T|} I(X; Y|Z_T). \tag{9}$$

We stress that the inner sum ranges over all *subsets* of S. For a description of set functions for which (8) and (9) are equivalent, see [20]. Note that we carefully distinguish between 3-way interaction information $II(X; Y; Z_S)$ with multivariate Z_S as the third component and $II(X; Y; Z_1, \ldots; Z_{|S|})$ being $(|S| + 2)$-way interaction information between X, Y and all components $Z_1, \ldots, Z_{|S|}$. Equality (8) can be restated, in view of $II(X; Y) = I(X; Y)$ and (6) as

$$I(X; Y|Z_S) = I(X; Y) + \sum_{i=1}^{|S|} \sum_{T \subseteq S, |T|=i} [II(X; Z_{t_1}; \ldots; Z_{t_i}|Y) - II(X; Z_{t_1}; \ldots; Z_{t_i})]. \tag{10}$$

Finally, note that it follows from (9) that $II(X; Y; Z_1, \ldots; Z_{|S|}) = 0$ provided X and Y are conditionally independent given any subvector Z_T of Z_S including $Z_\emptyset$. As we shall see in Section 4.1, Möbius expansion (8) is a natural starting point to introduce *CMI*-related criteria for feature selection.

Let us indicate that for any classifier $\hat{Y} = Y(X)$ of class variable Y for g classes, its unconditional probability of error $P(Y \neq \hat{Y})$ can be related to conditional entropy $H(Y|X)$ by means of Fano's inequality ([15,21])

$$H(Y|X) \leq 1 + P(Y \neq \hat{Y}) \log(g - 1),$$

which, using $I(X; Y) = H(Y) - H(Y|X)$, can be written as

$$P(Y \neq \hat{Y}) \geq \frac{H(Y) - I(X; Y) - 1}{\log(g - 1)}.$$

This shows that when $I(Y; X)$ decreases, i.e., Y and X become less dependent, the lower bound on probability of error of any classifier increases.

3. Feature Selection

In this section we discuss an objective of feature selection, the concept of Markov Blanket and its properties, as well as a greedy search for active features.

3.1. General Considerations: Characterisations of Markow Blanket

Suppose now that we consider class variable Y and p-dimensional discrete vector $(X_1, \ldots, X_p) =: (Z_1, \ldots, Z_p)$ of all predictors available. Our aim is to describe Y using those features among $X_1, \ldots, X_p$ which jointly influence Y. Note that this is a different task than selecting features which individually affect Y. Namely, we want to find out which

features can be discarded without loss of information about Y when the remaining features are taken into account. Thus, our aim is to check which features become redundant in the presence of other features. Moreover, we would like to investigate synergy between active features, that is a possible additional effect due to their simultaneous acting. Joint informativeness is, thus, crucial for feature selection. In this context, we define a minimal set of active predictors

Definition 5. *We define a minimal subset of active predictors as a minimal subset* $S^* \subset \{1, \ldots, p\} =: F$*, such that*

$$I(Y; X_{S^*}) = \max_{T \subseteq F} I(Y; X_T) = I(Y; X_F), \tag{11}$$

where X_T *denotes the subvector of* $(X_1, \ldots, X_p)$ *with indices in set* T*. Minimality is meant in the set-theoretic sense i.e.,* S^* *is minimal if there is no proper subset* $W \subset S$ *which satisfies (11).*

Note that the second equality in (11) is due to chain formula as it follows that $I(Y; X_T) \leq I(Y; X_{T'})$ for $T \subseteq T'$.

The problem of determining S^* is a feature selection problem stated in information-theoretic setting, as X_{S^*} may be considered as the minimal set describing adequately the overall dependence of Y on the available vector of predictors. The other possible way of defining a small subset of predictors which contain 'almost' all information about target Y is, for a given value of hyperparameter $\gamma > 0$, to consider the smallest subset S^*_γ, such that

$$I(Y; S^*_\gamma) \geq I(Y; X_F) - \gamma.$$

The other variant of the problem is its constrained version when a solution is sought for specific cardinality k of the feature set

$$I(Y; X_{S^*}) = \max_{T \subseteq F, |T| = k} I(Y; X_T), \tag{12}$$

which involves $\binom{|F|}{k}$ evaluations of mutual information. We also refer to the related bottleneck problem in which information in X is compressed to a random variable M satisfying a given compression condition making it easy to transmit, which has maximal mutual information with Y, see, e.g., [22].

Note that since $I(Y; X_S)$ is monotone in the second coordinate ($I(Y; X_S) \geq I(I(Y; X_T)$ for $T \subseteq S$), we have

$$I(Y; X_F) \geq \frac{1}{\binom{|F|}{k}} \sum_{T \subseteq F, |T| = k} I(Y; X_T),$$

thus the RHS can be used as a criterion when looking for S^* ([23]).

It is shown in [24] that when the distribution of $(X_1, \ldots, X_p)$ is determined by a subvector X_{S^*} corresponding to a certain $S^* \subseteq F$ and is parametrised by τ^*, finding a maximiser of $I(Y; X_T)$ is a first step of maximising the log-likelihood $\mathcal{L}(T, \tau)$ over T and τ under so called filter assumption stating that optimisations over T and over τ are independent. The problem of uniqueness of S^* satisfying (11) is important and sometimes disregarded as a problem in feature selection.

Remark 1. *Note that although (11) is trivially satisfied for* $S^* = F$ *it does not mean that the minimal set in the sense of inclusion is uniquely defined. The obvious example is constructed by taking any* Y, X*, such that* $I(Y; X) > 0$ *and letting* $X_1 = X$ *and* $X_2 = f(X_1)$ *where* f *is any* $1 - -1$ *transform which maps a set of values of* X *onto itself. In this case,* $I(Y; X_1) = I(Y; X_2) = I(Y; X_1, X_2)$*, thus both subsets* $\{X_1\}$ *and* $\{X_2\}$ *satisfy (11) and are minimal. Moreover, note that* $Y \perp\!\!\!\perp X_2 | X_1$*. However, uniqueness of* S^* *satisfying (11) holds for the case of continuous* X *and* Y *binary under strict positiveness assumptions stating that density* $p(x)$ *is positive almost everywhere with respect to Lebesgue measure and* $P_X(p(Y = 1|X) = 1/2) = 0$ *(see [25], p. 97).*

We discuss, now, properties of the set S^*. The first one is obtained by noticing that in view of the chain formula for S^* defined in (11) we have $I(Y; X_{S^{*c}}|X_{S^*}) = 0$, where T^c denotes complement of T in F. This is equivalent to stating that S^* is so called *Markov Blanket* of Y.

Definition 6. *Markov Blanket of target variable Y is the minimal subset $MB(Y) \subseteq F$, such that*

$$X_{MB(Y)^c} \perp\!\!\!\perp Y | MB(Y), \tag{13}$$

where $MB(Y)^c = \{1, \ldots, p\} \setminus MB(Y)$ (Minimal set satisfying (13) is also called Markov boundary).

Thus $MB(Y)$ shields Y from the rest of predictors in the sense that they become irrelevant once $MB(Y)$ is known. Again, $MB(Y)$ does not need to be uniquely defined. In order to discuss properties of $MB(Y)$ we introduce the concept of strong relevancy.

Definition 7. *X_i is strongly relevant feature provided*

$$I(Y; X_i|X_{F\setminus\{i\}}) > 0.$$

Note that the last property is equivalent to $Y \not\perp\!\!\!\perp X_i|X_{F\setminus\{i\}}$ and it can be restated as $P_{Y|X_F} \neq P_{Y|X_{F\setminus\{i\}}}$ [24]. Intuitively, strongly relevant features should be included in S^* as, after exclusion of such features, the dependence of Y on X_F is not adequately described. In the class of General Linear Models, under minimal conditions, features which are not strongly relevant are exactly those for which corresponding regression coefficient equals 0 (see [26], Proposition 2.2). The following facts concerning S^* have been formally proved:

Theorem 2. *Assume that Markov Blanket $MB(Y)$ exists. Then, we have:*

(i) *S^* coincides with Markov Blanket $MB(Y)$;*

(ii) *S^* satisfies*

$$\forall T \subset F, \;\; T \subseteq F \setminus S^* \iff Y \perp\!\!\!\perp X_T | X_{S^*\setminus T}; \tag{14}$$

(iii) *Every strongly relevant feature belongs to S^*.*

The first statement is justified above. Part (ii) is due to [27] and indicates that all subsets of F are partitioned into two parts: the first consisting of sets T disjoint with S^* for which $Y \perp\!\!\!\perp X_T|X_S^*$ holds and the remaining ones which are conditionally dependent with Y given $X_{S^*\setminus T}$.

We note that (iii) can be justified by the following simple reasoning: it is enough to show that if $j \notin S^*$ then X_j is not strongly relevant. Indeed, we have

$$\begin{aligned} & I(Y; X_{S^*}) = I(Y; X_F) = I(Y; X_{F\setminus\{j\}}) + I(Y; X_j|X_{F\setminus\{j\}}) \\ = \;& I(Y; X_{S^*}) + I(Y; X_{F\setminus\{\{j\}\cup S^*\}}|X_{S^*}) + I(Y; X_j|X_{F\setminus\{j\}}) \geq I(Y; X_{S^*}), \end{aligned}$$

where the second equality follows from the chain formula and the third from the fact that $j \notin S^*$ and the chain formula again. Thus, from the second equality it follows that $I(Y; X_j|X_{F\setminus\{j\}}) = 0$ whence X_j is not strongly relevant. It is not true in general that $MB(Y)$ consists only of strongly relevant features. Note that in the last example both X_1 and X_2 substituted for $MB(Y)$ satisfy (13), but neither of them is strongly relevant as $Y \perp\!\!\!\perp X_1|X_2$ and $Y \perp\!\!\!\perp X_2|X_1$. In the case when X is continuous and Y is binary, this can be proved for strictly positive distributions defined in Remark 1 (cf. Theorem 10 in [28]).

Ref. [27] introduces a concept of m-Markov Blanket S_m^* for $m \leq p$ for which equivalence in (14) is satisfied for any subset $T \subseteq F \setminus S_m^*$, such that $|T| \leq m$. As for $m = p$ the condition $|T| \leq m$ is vacuous, p-Markov Blanket of Y is $MB(Y)$.

3.2. Greedy Feature Selection

In the view of (i) of Theorem 2, finding S^* is equivalent to finding the minimal set satisfying

$$I(Y; X_{S^{*c}}|X_{S^*}) = 0.$$

This is a difficult task when p is large as it would involve checking this condition for all 2^p subsets of $\{1, \ldots, p\}$. The problem is frequently replaced by its greedy selection analogue: given S defined as the set of indices of features already selected, one determines

$$\arg\max_{j \in S^c}[I(X_{S\cup\{j\}}; Y) - I(X_S; Y)] = \arg\max_{j \in S^c} I(X_j; Y|X_S), \tag{15}$$

where the second equality follows from (4). The feature having the largest MI with Y is chosen first and the sequential search is stopped when for all $j \in S^c$ we have $I(X_j; Y|X_S) = 0$. Note that, in view of chain equality, the maximised criterion equals

$$I(X_j; Y|X_S) = I(X_j; Y) - I(X_j; X_S) + I(X_j; X_S|Y) = I(X_j; Y) + II(X_j; X_S; Y), \tag{16}$$

where $II(X_j, X_S, Y)$ is *three-way* interaction information of X_j, X_S, and Y. The first two terms of the first equality are called relevance and redundancy of X_j wrt Y and X_S, respectively. Note that in the maximisation process of $I(X_j; Y|X_S)$ over X_j, redundancy of X_j with respect to X_S may be outweighed by the magnitude of conditional information which X_j contains about X_S within classes: $I(X_j; X_S|Y)$. RHS of (16) involves $II(X_j, X_S, Y)$. In the next section, we show that $II(X_j, X_S, Y)$ is frequently replaced by algebraic expressions involving $II(X_j; X_{t_1}; \ldots; X_{t_k}; Y)$ where k does not exceed 2. This is motivated by (8) and allows for easier estimation of the expression.

4. Feature Selection Criteria Related to *CMI*

4.1. Criteria Based on Möbius Expansion

As estimation of conditional quantities, such as $I(Y; X|X_S)$ and its effective use for detecting conditional dependence in case when X_S is high-dimensional, requires large sample sizes (see, e.g., [29], Section 3.2), the common approach is to modify expressions, such as RHS of (10) to define analogues of CMI. Consider two approaches to achieve this aim. The first is based on truncation of the sum in (10) at a certain order, the second consists of weighing the corresponding terms and truncating them. In both cases, conditioning by random variable X_S is replaced with conditioning by low-dimensional sub-vectors of these vector, which drastically reduces the need for large sample sizes. This makes them easier to estimate and analyse. As a motivational example, consider the situation when X_S consists of $|S| = 10$ binary coordinates, uniformly distributed on $\{0, 1\}$ and the sample size $n = 1000$. Then, we expect around $1000/2^{10} \approx 1$ observation for each value $X_S = x_S$ of the conditioning variable X_S in $I(Y; X|X_S)$ whereas the respective number is around 500 for a single conditioning variable X_i appearing in the definition of, e.g., JMI in (21) below. The problem that low dimensionality of conditioning set facilitates estimation is recognised (see, e.g., [29]).

More specifically, we will consider below the following criteria based on truncation:

- Mutual Information Maximisation criterion MIM;
- Conditional Infomax Feature Selection criterion $CIFE$ of order two and three.

Additionally, the following criteria based on truncation and weighing of Möbius expansion will be considered:

- Generalised Information Criterion GIC;
- Joint Mutual Information criterion JMI of order two and three;
- Mutual Information Feature Selection criterion $MIFS$;
- Minimum-Redundancy Maximum-Relevance criterion $mRMR$.

The simplest approach is to consider only the first term of expansion (10) leading to *Mutual Information Maximisation* criterion $MIM(X) := I(X; Y)$ (cf. [30]). This completely

ignores dependence structure between Y and X_S, thus taking into account only feature relevance and disregarding its redundancy. However, it is useful and frequently applied for preliminary screening of predictors which are than subjected to more precise scrutiny. Actually, the name 'filters' is frequently used for exactly such criteria when MI is replaced by dependence measure between Y and X of user's choice.

Consideration of the first two summands of expansion (10) leads to *Conditional Infomax Feature Selection* ($CIFE$, [31], see also [24])

$$CIFE(X) = CIFE(X, Y|X_S) = I(X;Y) + \sum_{i \in S} II(Y;X;X_i), \tag{17}$$

which is also called Short Expansion of CMI of order 2 ($SECMI2$) in [29].

We stress that, in the definition of $CIFE$ and in definitions of the following criteria, the argument of the criterion is $X := X_i$ for $i \in S^c$ which is the variable over which maximisation is performed.

Analogously, taking into account the first three summands in (8) yield [32]

$$CIFE3(X) = I(X;Y) + \sum_{i \in S} II(Y;X;X_i) + \sum_{i<j, i,j \in S} II(Y;X;X_i;X_j), \tag{18}$$

which is also called, stressing its relation to CMI, $SECMI3$. Thus, for $|S| \leq 2$, $CIFE3$ in greedy selection rule yields the same results as (15).

Incorporating weights into (10) leads to the following definition of *Generalised Information Criterion* (GIC): for any $\beta, \gamma \in R^{|S|}$ we define

$$I_{\beta,\gamma}(X) = I(X;Y) + \sum_{k=1}^{|S|} \sum_{T=\{t_1,\ldots,t_k\} \subseteq S} [\gamma(k) II(X, X_{t_1};\ldots;X_{t_k})|Y) - \beta(k) II(X, X_{t_1};\ldots;X_{t_k})]), \tag{19}$$

where $\beta = (\beta(1),\ldots,\beta(k))$ and $\gamma = (\gamma(1),\ldots,\gamma(k))$. Usually $\beta(l) = \gamma(l) = 0$ for $l \geq l_0$, where l_0 is a predefined small integer. Letting $\beta(l) = \gamma(l) = 0$ for $l \geq 2$ one obtains criteria introduced in [24] parametrised by β and γ, where, abusing the notion slightly, $\beta := \beta(1)$ and $\gamma := \gamma(1)$:

$$J_{\beta,\gamma}(X) = I(X;Y) + \gamma \sum_{i=1}^{|S|} I(X;X_i|Y) - \beta \sum_{i=1}^{|S|} I(X;X_i). \tag{20}$$

For $\beta = \gamma = |S|^{-1}$ *Joint Mutual Information* criterion (JMI) is obtained [33]:

$$JMI(X) = I(Y;X) + \frac{1}{|S|} \sum_{i=1}^{|S|} (I(X;X_i|Y) - I(X,Z_i)) = I(Y;X) + \frac{1}{|S|} \sum_{k=1}^{|S|} II(Y;X;X_i). \tag{21}$$

Note that, in comparison to $SECMI2$ in (17), interaction terms are down-weighted by a factor $|S|^{-1}$.

Remembering that II is symmetric and writing $II(Y;X,Z_i) = I(Y;X|Z_i) - I(Y;X)$ one arrives at an useful form of JMI

$$JMI(X) = \frac{1}{|S|} \sum_{i=1}^{|S|} I(Y;X|X_i), \tag{22}$$

which, in particular, shows that $JMI = 0$ is equivalent to $Y \perp\!\!\!\perp X|X_i$ for any $i \in S$.

Remark 2. *Note that $JMI(X)$ is always non-negative, but it is not the case for $CIFE(X)$ (and neither for $CIFE3(X)$). Indeed, taking arbitrary X such that $I(Y,X_1) > 0$ and letting $X_2 = \cdots = X_p = X_1$ it is easy to check that w have for for $p > 2$:*

$$CIFE(X_1, Y|X_2,\ldots,X_p) = I(X_1;Y) + \sum_{j=2}^{p} (I(X_j;Y|X_1) - I(X_j;Y)) = -(p-2)I(X_1;Y) < 0$$

as $I(X_j;Y|X_1)=0$ *(cf. also [34]).*

Letting $\gamma=0$ and $\beta\in[0,1]$ one obtains *Mutual Information Feature Selection* (*MIFS*) criterion [35]

$$MIFS_\beta(X)=I(X;Y)-\beta\sum_{k=1}^{|S|}I(X;X_i),$$

with a special case $\beta=1/|S|$ called *Minimum-Redundancy Maximum-Relevance* criterion (mRMR) introduced in [36]. A normalised version of the *MIFS* criterion was considered in [37].

In [38], 3-way *JMI* has been introduced by starting from the equality

$$\sum_{\{i,j\}\subseteq S}I(Y;X,X_i,X_j)=\sum_{\{i,j\}\subseteq S}I(Y;X_i,X_j)+\sum_{\{i,j\}\subseteq S}I(Y;X|X_i,X_j).$$

Note that by dropping the first sum on RHS which does not depend on X (as the introduced criteria will be used to choose X) and scaling the second term by $2/(|S|(|S|-1)$ one obtains in view of (8)

$$\begin{aligned}JMI3(X) &= \frac{2}{S|(|S|-1)}\sum_{\{i,j\}\subseteq S}I(Y;X|X_i,X_j)\\ &= I(Y;X)+\frac{2}{|S|}\sum_{i\in S}II(Y;X;X_i)+\frac{2}{|S|(|S|-1)}\sum_{i<j}II(Y;X;X_i;X_j).\end{aligned}\quad(23)$$

This is generalised information criterion (20) with $\beta(1)=\gamma(1)=2/|S|$ and $\beta(1)=\gamma(1)=2/(|S|(|S|-1)$ and all other coefficients equal to 0.

The important criterion *Conditional Mutual Information Maximisation CMIM* using a different approach consisting of considering non-linear function of $I(Y;X|Z_i),i=1,\ldots,|S|$ has been proposed in [39]

$$CMIM(X)=\min_{j\in S}I(Y;X|X_j)=I(Y;X)-\max_{j\in S}[I(X;X_j)-I(X;X_j|Y)),\quad(24)$$

which should be maximised over $X:=X_i$ for $i\in S^c$. Thus, we look for the best surrogate of X_i among already chosen variables and the candidate having the worst the best surrogate of X is chosen. Generalisation of the rule based on *CMIM* is considered in [40].

4.2. Variational Approach

The other promising approach to approximate *MI* and *CMI* is based on construction of variational lower bounds of these quantities. The bounds obtained are then used as selection criteria. We discuss two approaches which are similar in nature. The first one is based on Donsker–Varadhan inequality which states that [41]

$$I(Y;X)\geq\sup_{f\in\mathcal{F}}[E_{P_{X,Y}}f(X,Y)-\log E_{P_X\otimes P_Y}e^{f(X,Y)}],\quad(25)$$

and inequality becomes equality when family $\mathcal{F}$ contains the logarithm of the ratio of densities of $P_{X,Y}$ and $P_X\otimes P_Y$, namely $h(x,y)=\log(f_{XY}(x,y)/f_X(x)f_Y(y))$. Indeed, note that by plugging $h(x,y)+C$ into the expression on the RHS of (25) we obtain equality. Estimation of $I(X;Y)$ based on the Donsker–Varadhan formula has been proposed in [42] where neural network is applied for $\mathcal{F}$ resulting in MINE method. The approach has been further pursued in [43,44]. Other lower bounds, such as Nguyen–Wainwright–Jordan bound [45] can also be used. Note that in order to evaluate the expected value under independence in (25) one uses permutations of the original sample which consists in permuting X values while keeping the values of Y at their original places (see Section 7). For feature selection, one can either directly evaluate $I(Y;X_{S\cup\{j\}})$ or $I(Y;X_j|X_S)$. For the second approach resampling methods which would ensure conditional independence for

generated data are needed. Note that the efficiency of such methods depends on flexibility of function class $\mathcal{F}$ over which the bound in (25) is optimised. The empirical evidence suggest that relatively simple neural nets are sufficiently flexible to yield satisfactory estimators of $I(Y;X)$. They are also preferable, as large networks will increase the variability of the estimators.

The second bound, similar in flavour to (25), is due to [46]. It is noticed there that for arbitrary density $q(x,y)$ we have

$$\begin{aligned} I(Y;X) &= H(Y) - H(Y|X) = H(Y) + E_{P_X} E_{P_{Y|X}} \log p(Y|X) \\ &\geq H(Y) + E_{P_X} E_{P_{Y|X}} \log q(Y|X) = E_{P_{X,Y}} \log \frac{q(X|Y)}{p(Y)}, \end{aligned}$$

where the inequality is due to $D(p(Y|x)||q(Y|x)) \geq 0$. Assuming that $q(y) = p(y)$ in view of Bayes theorem it is now sufficient to specify marginal density $q(x|y)$ to obtain a lower bound on $I(Y;X)$. In [46] $q(x|y)$ are considered which either satisfy naive Bayes assumption or more general assumption that $q(x_t|y,x_1,\ldots,x_{t-1})$ is geometric mean of $q(x_t|y,x_i)$ for $i = 1,\ldots,t-1$. Other proposals, using specific tree-representation for $q(x|y)$ are possible (see, e.g., [47]).

5. Interplay between *CMI* and *CMI*-Related Criteria

The following result states conditions under which CMI and one of the introduced criteria coincide. As before, X denotes a feature from $X_{F\setminus S}$ considered as a candidate to be added to $\{X_s, s \in S\}$.

Theorem 3.

(i) *Assume that all features are conditionally independent given class (naive Bayes assumption): and features in S are conditionally independent given any not chosen feature (i.e., belonging to S^c). Then, for $X \in X_{F\setminus S}$, $I(Y;X|X_S)$ differs from $MIFS_1(X) = I(Y;X) - \sum_{i\in S} I(X;X_i)$ by a factor which does not depend on X.*

(ii) *Assume that all k-way interaction informations $II(Y;X;X_{t_1};\ldots;X_{t_l}) = 0$ for $k > 3$ ($l > 1$). Then, $I(Y;X|X_S) = CIFE(Y,X|X_S)$ for any $X \in X_{F\setminus S}$.*

(iii) *Ref. [24] If $X_i, i \in S$ are conditionally independent given X, for $X \in X_{F\setminus S}$ and additionally they are conditionally independent given X and Y, then $I(Y;X|X_S)$ differs from $CIFE(Y,X|X_S)$ by a factor which does not depend on X.*

(iv) *Ref. [48] Assume that for any $i \in S$ we have $X \perp\!\!\!\perp X_{S\setminus\{i\}}|X_i$ and additionally $X \perp\!\!\!\perp X_{S\setminus\{i\}}|X_i, Y$. Then, $I(Y;X|X_S) = JMI(Y;X|X_S)$.*

For completeness, the proof is included in the Appendix A. Note that the conditions imposed by (i) are very stringent. There are no known probabilistic vectors satisfying $I(Y;X;X_{t_1},\ldots;X_{t_l}) = 0$ for $l > 1$ apart from simple situations, such as $X_S \perp\!\!\!\perp (X,Y)$, which obviously does not hold when X_S is the set of chosen predictors.

We remark that under conditions of (i) $MIFS_1(X) = CIFE(X)$, as due to the naive Bayes assumption $\sum_{i\in S} I(X_i;X|Y) = 0$. Additionally, it follows from (21) that $JMI(X) = mRMR(X)$ provided that the naive Bayes condition holds, thus in order to have $JMI(X) = CMI(X) = mRMR(X)$ under rather strong assumptions of (iv), we have to assume additionally naive Bayes condition. This indicates that the last equality can hold only under restrictive conditions and is not true in general (see [24], Section 4.1).

Assumptions in (iii) and (iv) are not compatible as they lead to different forms of criteria equivalent to CMI criterion. It is argued in [46] that the only plausible graph representation of probability structure for which conditions of (iii) is graph with edges $E = \{(Y,X)\} \cup \{(X,X_i)\}_{i\in S}\}$. In this case, due to data-processing inequality, we have $I(Y;X) \geq I(Y;X_S)$. This means, however, that X should have been chosen *before* any features from S.

Additional results of similar flavours to Theorem 3 are discussed in [38,49].

It follows from preceding discussion that criteria introduced in Section 4.1 are formally introduced by truncation (or truncation and weighing) of terms in Möbius expansion. Their analytical properties concerning how well they approximate CMI have yet to be established.

6. Asymptotic Distributions of Information-Theoretic Measures

Sequential feature selection for predicting target variable Y typically involves checking whether a new candidate feature X is a valuable addition to the set of already chosen features X_S. This is usually based on testing whether null hypothesis $H_0 : X \perp\!\!\!\perp Y|X_S$ holds and its rejection is interpreted as an indication that X carries an additional information about Y to that provided by X_S. To this end, $\widehat{I}(Y;X|X_S)$ or its modified versions are used. The usual strategy following statistical testing approach is to derive asymptotic distribution of $\widehat{I}(Y;X|X_S)$ or its modifications described in Section 4 under the null. This distribution is used as a benchmark for which value of the statistic for the sample under consideration is compared to obtain the asymptotic p-value of the test. In the following, we describe the asymptotic distribution of $\widehat{CMI}$ and $\widehat{CMI}$-related criteria under H_0. A competing approach to approximate the distribution of the considered statistic under H_0 based on resampling is described in Section 7.

6.1. Asymptotic Distribution of $\widehat{CMI}$

We assume that X, Y, X_S take I, J, K possible values, respectively. As X_S is a $|S|$-dimensional vector, K is the number of all possible combinations of values of its coordinates. We let $p(x,y,x_S) = P(X = x, Y = y, X_S = x_S)$ and consider the case when all probabilities $p(x,y,x_S)$ are positive. It is assumed throughout that the estimation of $I(Y;X|X_S)$ and related quantities is based on n independent identically distributed (iid) samples from the distribution of (X,Y,X_S). Construction of estimators of CMI relies on plugging-in frequencies in place of unknown probabilities, e.g.,

$$\hat{I}(Y;X|X_S) = \sum_{x,y,x_S} \hat{p}(x,y,x_S) \log \frac{\hat{p}(x,y|x_S)}{\hat{p}(x|x_S)\hat{p}(y|x_S)} = \sum_{x,y,x_S} \hat{p}(x,y,x_S) \log \frac{\hat{p}(x,y,x_S)\hat{p}(x_S)}{\hat{p}(x,x_S)\hat{p}(y,x_S)}, \tag{26}$$

where $\hat{p}(x,y,x_S) = n(x,y,x_S)/n$ and $n(x,y,x_S)$ is a number of samples equal to (x,y,x_S). We will consider only frequencies as estimators of discrete probabilities. Other estimators exist, e.g., regularised versions of sample frequencies, for which regularisation reflects a level of departure from conditional independence, see [38,50]. The following known result is frequently used in dependence analysis (compare [51], see also [52]).

Theorem 4.

(i) Assume that $I(Y;X|X_S) \neq 0$. Then, we have

$$n^{1/2}(\widehat{I}(Y;X|X_S) - I(Y;X|X_S)) \xrightarrow{d} N(0, \sigma^2_{\widehat{CMI}}), \tag{27}$$

where

$$\sigma^2_{\widehat{CMI}} = \sum_{x,y,x_S} p(x,y,x_S) \log^2 \frac{p(x,y,x_S)p(x_S)}{p(x,x_S)p(y,x_S)} - I^2(X,Y|X_S) = \mathrm{Var}\left(\log \frac{p(X,Y,X_S)p(X_S)}{p(X,X_S)p(Y,X_S)}\right)$$

and $\sigma^2_{\widehat{CMI}} > 0$.

(ii) Assume that $I(Y;X|X_S) = 0$. Then,

$$2n\widehat{I}(Y;X|X_S) \xrightarrow{d} \chi^2_d, \tag{28}$$

where $d = (I-1)(J-1)K$.

The frequently applied test of conditional independence $X \perp\!\!\!\perp Y|X_S$ is based on the above useful fact (ii) that under independence asymptotic distribution $\hat{I}(Y;X|X_S)$ does not depend on the distribution of (X,Y,X_S) and is chi-square with the known number of degrees of freedom. On the other hand, when the conditional independence does not hold, the limiting distribution of $\hat{I}(Y;X|Z) - I(Y;X|X_S)$ is normal with the variance depending on the underlying probability distribution P_{XYZ}. We stress that speeds of convergence of $\hat{I}(Y;X|X_S)$ to $I(Y;X|X_S)$ are different in both cases: they equal n^{-1} in the first case and $n^{-1/2}$ in the second.

The test based on CMI is a popular tool in dependence analysis, in particular for Markov Blanket discovery (see Section 8). It has different names among which G^2 test is the most popular (see, e.g., [53]). Additionally, in the literature X^2 denotes the second order approximation of $\widehat{CMI}$, which turns out to be the conditional chi-square test and has the same asymptotic distribution as $\hat{I}(Y;X|X_S)$.

6.2. Asymptotic Distribution of Modified Criteria

Asymptotic distribution of the modified criteria related to $\widehat{CMI}$ can be also derived. Let $\widehat{J^{\beta,\gamma}}(X,Y|Z)$ be a plug in-version of $J^{\beta,\gamma}$ defined in (20). Moreover, let $p = (p(x,y,x_s))$ be a vector of probabilities of dimension $M = I \times J \times K$, $\hat{p}$ corresponding vector of fractions and $f : [0,1]^M \to R$ be a function which represents $J^{\beta,\gamma}$ as a function of p, i.e., $J^{\beta,\gamma} = f(p)$. For example for JMI criterion (see (22)) the corresponding function is defined as

$$f_{JMI}(p) := \sum_{x,y,x_S} p(x,y,x_S) \frac{1}{|S|} \sum_{s \in S} \ln \frac{p(x,y,x_s)p(x_s)}{p(x,x_s)p(y,x_s)},$$

where x_s denotes coordinate of x_S. We state here the result on asymptotic distribution of $\widehat{J^{\beta,\gamma}}(X,Y|X_S)$ proved in [29]. Let $\Sigma_{x,y,x_S}^{x',y',x'_S}$ denote an element of matrix Σ with row index x,y,x_S and column index x',y',x'_S.

Theorem 5.

(i) We have

$$n^{1/2}(\widehat{J^{\beta,\gamma}}(X,Y|X_S) - J^{\beta,\gamma}(X,Y|X_S)) \xrightarrow{d} N(0,\sigma_f^2), \tag{29}$$

where $\sigma_f^2 = Df(p)^T \Sigma Df(p) = \mathrm{Var}(Df(p)^T \hat{p})$ *and* $\Sigma = n\Sigma_{\hat{p}}$ *is a matrix consisting of elements* $\Sigma_{x,y,x_S}^{x',y',x'_S} = p(x',y',x'_S)(I(x=x',y=y',x_s=x'_S) - p(x,y,x_S))/n$.

(ii) If $\sigma_f^2 = 0$ *then*

$$2n(\widehat{J^{\beta,\gamma}}(X,Y|X_S) - J^{\beta,\gamma}(X,Y|X_S)) \xrightarrow{d} V^T H V, \tag{30}$$

where V *follows* $N(0,\Sigma)$ *distribution, and* $H = D^2 f(p)$ *is a Hessian of* f.

In particular, we have the following result for $\widehat{JMI}$ [54]:

Corollary 1. *Let* Y *be binary.*

(i) If $\sigma_{\widehat{JMI}}^2 \neq 0$ *then*

$$n^{1/2}(\widehat{JMI} - JMI) \xrightarrow{d} N(0,\sigma_{\widehat{JMI}}^2,$$

where

$$\sigma_{\widehat{JMI}}^2 = \sum_{x,y,x_S} p(x,y,x_S) \left(\frac{1}{|S|} \sum_{s \in S} \ln \frac{p(x,y,x_s)p(x_s)}{p(x,x_s)p(y,x_s)} \right)^2 - (JMI)^2 \tag{31}$$

(ii) *We* $\sigma^2_{\widehat{JMI}} = 0 \iff JMI = 0$. *In this case,*

$$2n\widehat{JMI} \xrightarrow{d} V^T H V = \sum_{i=1}^{M} \lambda_i Z_i^2,$$

where V *and* H *are defined in Theorem 2,* Z_i *are iid* $N(0,1)$ *and* λ_i *are eigenvalues of the matrix* $W = \Sigma H$ *which has the following elements*

$$W^{x',y',x'_S}_{x,y,x_S} = \frac{1}{|S|}\sum_{s=1}^{|S|}\left[\frac{I(x_s = x'_s)}{p(x_s)} - \frac{I(x = x', x_s = x'_s)}{p(x, x_s)} - \frac{I(y = y', x_s = x'_s)}{p(y, x_s)} + \frac{I(x = x', y = y', x_s = x'_s)}{p(x, y, x_s)}\right].$$

It follows from Theorem 1 that if for any $i \in S$ we have that $I(Y; X|X_i) = 0$, i.e., Y and X are conditionally independent given X_i, then asymptotic distribution is that of quadratic form specified in (ii), otherwise the distribution is normal.

The main advantage of using modified criteria instead of CMI is that their estimation does not require as large samples as for CMI itself. Note that for modifications of order k, conditioning involves k-dimensional strata, whereas for CMI p-dimensional strata are considered. However, modified criteria considered in Section 4.1 suffer from the fact that under hypothesis of conditional independence $X \perp\!\!\!\perp Y|Z$ the asymptotic distribution of empirical criterion is not uniquely determined and has to be estimated from the sample. As in the case of $\widehat{JMI}$, the asymptotic distribution can be either normal (when $X \not\perp\!\!\!\perp Y|X_j$ for at least one $j \in S$ or coincides with distribution of the quadratic form. Which type of asymptotic distribution is valid can be decided using resampling schemes shortly discussed in Section 7. The chosen distribution can used as a benchmark to test the conditional independence hypothesis (see, e.g., [29,55]). Alternatively, testing of $H_0 : I(Y; X|X_S) = 0$ can be replaced by testing $\tilde{H}_0 : I(Y; X|X_i) = 0$ for any $i \in S$ at each stage of forward feature selection procedure and the benchmark distribution can be obtained by approximating eigenvalues of M specified in (ii) (see [54]) or using scaled and shifted χ squared distribution $\alpha + \beta\chi^2_d$, where α, β, d are estimated from the sample [56]. Note that, although H_0 and $\tilde{H}_0$ are not equivalent, in the case of faithful distributions (See Section 9.2 for definition of faithfulness) $\tilde{H}_0$ implies H_0, as conditional independence is inherited by conditioning supersets in such a case.

7. Resampling Schemes

When using the conditional independence test, which the test statistic is used for, what is very common, the exact or even the asymptotic distribution under conditional independence is not known, the usual practice is to use conditional randomisation (CR) and resampling schemes to approximate this distribution and use the resulting approximation as the benchmark distribution, as described in the beginning of Section 6. Analogously as before, comparison of the value of the statistic based on the observed sample with the benchmark distribution is used to calculate the CR or resampling p-value. It can be also used as alternative way of assessing the distribution of the statistic even when its approximate distribution is known. We briefly discuss these procedures, first for a discrete X_S, then for a continuous case.

- *Conditional Randomisation* (CR).In the case of the CR approach we assume that the probability mass function $p(x|X_S = x_S)$ or density of X given $X_S = x_S$ is known. Then, given a sample $(\mathbf{X}, \mathbf{Y}, \mathbf{X}_S) := (X_i, Y_i, X_{S,i})_{i=1}^n$ one generates a CR sample $(\mathbf{X}^*, \mathbf{Y}, \mathbf{X}_S) = (X_i^*, Y_i, X_{S,i})_{i=1}^n$, when X_i^* are independently sampled from p.m.f. or density $p(x|X_S = x_{S,i})$ for $i = 1, \ldots, n$. Let $T(\mathbf{X}, \mathbf{Y}, \mathbf{X}_S)$ be a test statistic which we would like to use for testing CI and consider M generated CR samples. We define the permutation p-value as

$$p_{CR} = \frac{\#\{k = 1, \ldots, M : T(\mathbf{X}_k^*, \mathbf{Y}, \mathbf{X}_S) \geq T(\mathbf{X}, \mathbf{Y}, \mathbf{X}_S)\} + 1}{M + 1}. \tag{32}$$

It follows that p_{CR} is a valid p-value, i.e., conditionally on $((\mathbf{Y}, \mathbf{X}_S)$, i.e., its distribution is uniform on the set $\{1/M, 2/M, \ldots, M/(M+1)\}$ under CI and, thus,

$$P(p_{CR} \leq \alpha | \mathbf{Y}, \mathbf{Z}) \leq \alpha.$$

This follows from the following simple fact. Suppose that CI holds and $\mathbf{X}^*$ is such that $P_{\mathbf{X}^*|\mathbf{X}_S} = P_{\mathbf{X}|\mathbf{X}_S}$. Then, given $\mathbf{Y}, \mathbf{X}_S$, when CI holds, we have the following equality in distribution

$$T(\mathbf{X}, \mathbf{Y}, \mathbf{X}_S) \stackrel{d}{=} T(\mathbf{X}^*, \mathbf{Y}, \mathbf{X}_S);$$

(see, e.g., [57]). For recent improvements of the CR method, random permutations are appropriately chosen and considered instead of choosing them at random, see [58]. In the case when $p(x|X_S = x_S)$ is unknown, CR sampling is replaced by Conditional Permutations or Bootstrap X method.

- *Conditional permutation (CP)*. Conditional permutation method is similar to *CR* method and differs in that for value x_S taken by elements of $\mathbf{X}_S$ we consider the strata of the sample corresponding to this value, namely

$$P_i = \{j : (X_j, Y_j, X_{S,j}) : X_{S,j} = x_{S,i}\}.$$

The CP sample is obtained from the original sample by replacing $(X_j, Y_j, X_{S,j})$ for $j \in P_i$ by $(X_{\pi^i(j)}, Y_j, X_{S,j})$, where π^i is a randomly chosen permutation of P_i. Thus, on every strata $X_S = x_s$ we randomly permute values of corresponding X independently of values of Y. Once M of CP samples are obtained independently in this fashion, we calculate p-value p_{PC} based on them analogously as before.

- *Conditional Bootstrap X (CB.X)*. Instead of permuting values of X on each strata, we draw a bootstrap sample from X observations, that is why we sample them with replacement as many times as is the size of the strata. The remaining steps are as previously described.

We discuss now the resampling for continuous predictors. Feature selection for this case is covered shortly in Section 9. We remark here that the aim, methodology of solutions, and criteria considered are very similar, the main differences consist of technical problems of adequate estimation of information-theoretic measures in the continuous case.

In the case of X_S being continuous, CR methods work as stated, however the situation is more complicated for *CP* and *CB.X* method as those depend on transforming the strata of the sample, which degenerate to single observation points in this case. One possible solution is to sample from some estimate $\hat{p}(x|X_S = x_S)$, e.g., kernel estimate, however this requires a large number of observations on each strata. The following proposal of constructing pseudo sample distributions of which is close to P_{X,Y,X_S} under CI has been suggested by [59] and consists in using observations having close values to $X_S = x_S$. More specifically, consider observation $(x_i, y_i, x_{S,i})$ and find k^{th} nearest observation to $x_{S,i}$ in X_S space, say, $x_{S,j}$ with corresponding triple $(x_j, y_j, x_{S,j})$. Then, observation of $(x_i, y_i, x_{S,i})$ is replaced in the resampling sample by $(x_j, y_i, x_{S,i})$.

Another technique for data augmentation, applicable in both discrete and continuous cases, is *knock-off* construction which we now describe. Consider the regression problem involving vector $X = (X_1, \ldots, X_p)$ of features and response Y. Vector $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ is the knock-off vector for X in this regression problem, if $Y \perp\!\!\!\perp \tilde{X}$ and, moreover, for any $S \subseteq \{1, \ldots, p\}$ we have the following equality in distribution [57]:

$$(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p)_{swap(S)} \stackrel{d}{=} (X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p,)$$

where $swap(S)$ means swapping entries X_j and $\tilde{X}_j$ for any $j \in S$. Thus, in the above sense $\tilde{X}_1, \ldots, \tilde{X}_p$ are interchangeable with $X_1, \ldots, X_p$ and independent of Y at the same time.

The construction of knock-offs is complicated in general, feasible only in special cases as of now, such as the Gaussian case, and requires knowledge of distribution of P_X. Thus,

even more information is needed than in the case of CR approach. However, the gains are considerable, as for natural classes of statistics measuring influence of predictors on the response one can compare the performance of $X_1, \ldots, X_p$ with that of their knock-offs and on this basis decide which ones are not strongly relevant. Intuitively, when the performance of X_j measured, e.g., by an absolute value of estimated regression coefficient, is comparable to that of its knock-off, then such a variable should be discarded. This approach yields a bound on FDR of strongly relevant features (Theorem 3 in [26]). Thus, in cases discussed in Section 3, when the set of strongly relevant features is exactly equal $MB(Y)$, this approach yields a bound on FDR of its recovery. We stress that, in this way, one can analyse properties of the whole procedure of MB determination, and not only that of an individual steps in the algorithm. An additional advantage is that, unlike for the resampling schemes above, only one sample of knock-offs needs to be generated.

8. Markov Blanket Discovery Algorithms

We discuss now Markov Blanket discovery algorithms. Their aim is to solve a feature selection problem posed in Section 3 (see Equation (11)) applying CMI or CMI approximations introduced in Section 4 for which theoretical guarantees are discussed in Section 5. Such algorithms are used as building blocks for conditional independence tests, based either on asymptotic distributions of test statistics discussed in Section 6 or approximations of their distributions based on resampling covered in Section 7.

We discuss three representative examples of Markov Blanket discovery algorithms. Other examples include [60] and algorithms using assumed faithful Directed Acyclic Graph (DAG) representation of the underlying probability structure (for DAG fathfulness of probability distribution based on d-separation (see e.g., [4]), such as HITON-MB [61], MMMB [62], IPC-MB [63], and STMB [64].

- The GS (Grow and Shrink) algorithm [65]. It consists of two phases. Specific ordering of variables in F is considered, e.g., variables may be ordered according to value of $I(Y;X)$ and, then, the variable having the largest value of the mutual information is the first variable chosen. Denoting by S the current set of chosen variables, we pick the first variable (in the considered ordering) which depends on Y given X_S ($I(Y;X|X_S) > 0$) and we add it to S, then repeat the step. When there is no longer such a variable among candidates, the first phase is terminated. We call the resulting chosen set S^*. In the second phase, we remove from S^*, again using the considered ordering, any variable X_j which is not strongly relevant with respect to the current S^*, i.e., such that $X_j \perp\!\!\!\perp Y|X_{S^*\setminus\{j\}}$ ($I(Y;X_j|X_{S^*\setminus\{j\}}) = 0$) and we let $S^* := S^* \setminus \{j\}$.
- The IAMB (Incremental Association Markov Blanket) [66]. The algorithm is similar to GS with one important difference in the first step. Namely, it disregards initial ordering and S is augmented by the most plausible candidate, that is the variable realising $\max_{i\in S^c} I(X_i, Y|X_S)$, provided it is not conditionally independent from Y given X_S.
- $GS^{(m)}$ ([27], where $m \leq p$. $GS^{(m)}$ differs from GS in the growing phase only. Namely, instead of an *individual* variables with indices in $F \setminus S$ being considered as possible candidates, all *subsets* T of size not exceeding m are taken into account and the check is performed whether there are conditionally dependent on Y given current S. If this holds $S := S \cup T$.

It is proved in [27] that $GS^{(m)}$ algorithm yields the m-Markov Blanket (see Section 3 for definition of m-Markov Blanket) of Y. Thus, for $m = p$ we have that the output of $GS^{(p)}$ is $MB(Y)$. However, since, in the growing phase, all subsets of $F \setminus S$ of size not exceeding m have to be checked which is computationally intensive, in practice k subsets for k large enough are checked at every step.

Note that for such results we assume that condition $Y \perp\!\!\!\perp X_{T\setminus S}|X_S$ or $Y \perp\!\!\!\perp X|X_S$ can be verified. In practice, it is impossible to check conditional independence of X and Y given the set of variables already chosen without error and this has to be replaced with the appropriate test discussed in Section 6. Obviously, as such a test has to be performed at each

step, one does not control probability of including false discoveries. False discoveries are dealt with in the second phase, but no formal results exist concerning how likely recovering $MB(Y)$ is in the case of practical algorithm. However, see [67] for the results concerning the PC algorithm in the Gaussian case and knock-off construction discussed in Section 7.

Another possibility is to omit phase two and stop early phase one, applying stopping rules devised in multiple testing problems to control Family-Wise Error Rate or False Discovery Rate (see, e.g., [68] where a stopping rule using the Holm procedure has been applied for CIFE criterion). These methods work promisingly in practice, however, due to the fact that the set S augmented at each step is data-dependent, also in this case formal results on the consistency of such procedures are not available.

There is no definitive study of empirical performance of the presented criteria and/or selection procedures; note that any criterion considered can be used for any feature selection method described above, thus creating a large number of filter methods. Moreover, those can be evaluated using their classification performance, as well as for synthetic datasets, according to their ability to choose active predictors. Thus, we discuss only the reports on the simplest greedy procedures consisting on the application of discussed criteria on synthetic datasets with a fixed ahead number of chosen features. Authors of [24] discuss, among other things, results for datasets coming from NIPS Feature Selection Data Challange, Gisette, and Madelon (procedures stopped at 200 variables in the first case and 20 in the second). Two interesting points arise from the study: strong performance of JMI, which was the second best and co-winner in those cases, with respect to balanced accuracy (BA) and the number of feature chosen, with a strong performance of $CMIM$ (winner in the case of Gisette and together with $CIFE$ co-winner in the case of Madelon). The overall strong performance of $CIFE$ was also confirmed in [68,69]. The second important observation is the failure of performance of CMI, which due to scarcity of data, failed to detect conditionally dependent variables very early. This is due to the fact that for a large conditioning set the test becomes very conservative (see also [70] for discussion of this property).

9. Case of Continuous Distribution $P_{X,Y,Z}$

In the following, we shortly discuss conditional independence testing for continuous distributions. The main motivation to include a continuous case in this review, devoted mainly to discrete case, is to underline the strong similarities between these two cases. In particular, we discuss below that all information-theoretic tools defined for the discrete case have their analogues for continuous distributions. Moreover, we note that the selection methods presented till now do not depend on continuity of the underlying measure, once a specific test for conditional independence $X \perp\!\!\!\perp Y|Z$ which takes this into account is used. Thus, the differences between two cases are mostly due to the fact that estimation of information-theoretic measures are much more difficult for continuous distributions. Moreover, their asymptotic behaviour under CI is not known, thus making construction of corresponding tests difficult. On the other hand, for Gaussian case significant simplifications exists due to the existence of closed formula for CMI discussed in Section 9.2.

9.1. General Considerations

Returning to notation of Section 2, we discuss conditional independent testing in a continuous case. In this context, it is enlightening to mention the result in [71], where it is shown that in the continuous case conditional independence testing is a hard problem in the following sense. For any test on a natural family of $P_{X,Y,Z}$ defined in [71] satisfying $X \perp\!\!\!\perp Y|Z$, which *uniformly* achieves type I error asymptotically not larger than a prescribed significance level α its power for *any* alternative will be also no larger than α. Thus, such a test is useless for discriminating between CI and Conditional Dependence.

The indices defined in Section 2 have their natural analogues in the continuous case; namely probabilities $P(X = x, Y = y, Z = z)$ are replaced by density functions $p(x, y, z)$ and summations by integrals (for details see [15]). What is important for development

here is that Conditional Independence $X \perp\!\!\!\perp Y|Z$ in the continuous case is also equivalent to conditional Mutual Information $I(X;Y|Z)$ being 0, as in the discrete case. Thus, the problem boils down to an accurate estimation of $I(X;Y|Z)$ and corresponding testing of CI using constructed estimator as a test statistic. This, however, is also a much harder task than in a discrete case as plug-in estimators of entropy, and conditional and unconditional information when, e.g., kernel estimators of densities (see, e.g., [6], Chapter 6) are plugged-in are unstable and work satisfactorily only when very large samples are available. Much better results are obtained using the variational approach described in Section 4.2. Additionally, some estimators based on the nearest neighbour idea, such as the Kozachenko–Leonenko estimator of entropy [72,73] and its refinements, such as the Kraskov et al. estimator [74] work better than straightforward plug-in estimator. Then, estimators of MI are constructed using (1). Additionally, the simple approach based on discretising underlying variables is frequently used. There are two problems related to this approach. The first one is a loss of information due to this operation. The second, related one, is a difficulty of choosing an appropriate bin size, especially in many dimensions. Too large bin size may result in hiding interesting characteristics of predictors' distributions and, consequently, a loss of predictive power.

We mention two other approaches for the continuous case. One which is frequently used is *kernel based approach* which relies on the following fact [75]:

$$X \perp\!\!\!\perp Y|Z \iff Ef(X,Z)h(Y,Z) = 0 \text{ for any } f \in \mathcal{F}_{X|Z}, h \in \mathcal{F}_{Y|Z},$$

where $\mathcal{F}_{X|Z} = \{h(Y,Z) : h(Y,Z) = h_0(Y) - E(h_0(Y)|Z), h_0 \in L^2_Y\}$. Thus, although conditional independence is not equivalent to conditional covariance being zero, this is true when transformations of vectors (X,Z) and (Y,Z) are allowed. In view of this result it turns out that its possible to find rich enough function spaces, such that the condition that the generalized conditional covariance of transformed X and Y being equal to 0 is equivalent to conditional independence. Moreover, it is the function space that one can consider to appropriately define Reproducing Kernel Hilbert Spaces (RKHS). This, conceptually, is much more involved than checking $I(X;Y|Z)$ being 0, save for the technical problems of testing this condition in the continuous case.

We also mention in this context the second approach, called distillation method [76] which relies on ideas similar to the construction of partial residual plot and resampling.

9.2. Gaussian Case

We review the case when (Y,X) is Gaussian, where $X = (X_1, \ldots, X_p)$. This case offers significant simplifications, as the quantities on which feature selection is based can be explicitly calculated. Namely, direct calculation yields [15]:

$$H(X) = \frac{1}{2}\log|\Sigma_X| + \frac{p}{2}\log(2\pi e),$$

if $X \sim N(\mu_X, \Sigma_X)$. If $(X,Y) \sim N(\mu_{X,Y}, \Sigma_{X,Y})$ it follows from (2) that

$$I(X;Y) = \frac{1}{2}\log\frac{|\Sigma_{X,Y}|}{|\Sigma_X| \times |\Sigma_Y|},$$

where $|\Sigma|$ stands for determinant of matrix Σ. In particular for bivariate normal case when $\rho(X,Y) = \rho$ we have $I(Y;X_1) = -(\log(1-\rho^2))/2$. Whence, in the case of the conditional distribution of (Y,X), given X_S, where $X_S \subseteq \{2,\ldots,p\}$ this yields

$$I(Y;X_1|X_S) = -\frac{1}{2}\log(1-\rho^2_{par}),$$

where ρ_{par} is partial correlation coefficient between Y and X_1 given X_S. Thus, in Gaussian case information-based dependence indices can be expressed in terms of either correlation or partial correlation coefficients.

In order to present the PC algorithm, we first introduce the concept of the faithfulness of the distribution which plays an important role in its construction (see, e.g., [67]). Let $X_0 := Y$ and consider undirected graph $G = (V, E)$, where vertices $V = \{0, 1, \ldots, p\}$ correspond to variables $(Y, X_1, \ldots, X_p)$ and $E \subset V \times V$, such that $(j, k) \in E \iff (k, j) \in E$ is the set of edges.

Definition 8. *Distribution $P_{Y,X}$ is faithful to graph G when for any triple $A, B, C \subset V$ we have*

$$A \text{ and } B \text{ are separated by } C \iff X_A \perp\!\!\!\perp X_B | X_C, \tag{33}$$

when separation by C means that every path joining elements of A and B has to pass through C. (Implication from LHS to RHS in (33) is called global Markov property)

In the case when $(Y, X_1, \ldots, X_p)$ is Gaussian and there exists faithful representation of its distribution, PC algorithm [77] reconstructs its dependence structure with probabilistic guarantees. The main tool is clever usage of one of the consequences of faithfulness, namely

$$X_1 \perp\!\!\!\perp X_2 | X_{C_1} \Rightarrow X_1 \perp\!\!\!\perp X_2 | X_{C_2} \text{ for any } C_2 \supseteq C_1. \tag{34}$$

The algorithm starts from the fully connected graph and removes edges between any vertices which are marginally uncorrelated (and, thus, in view of (34) and Gaussianity, conditionally independent given any subset of variables). Then, it proceeds to remove edges between vertices which correspond to variables which are conditionally independent given a subset of their neighbours of size l, where $l = 1, 2, \ldots$ is increased stepwise. The conditional independence is tested using an empirical partial correlation coefficient $\hat{\rho}_{par}$ and the property that its Fisher transform $log[(1 + \hat{\rho}_{par})/(1 - \hat{\rho}_{par})]$ is approximately $N(0, 1)$ normal under CI (details in [67], Section 13.7).

The case when Y is discrete and $P_{X|Y=y_i}$ are normal is more complicated, as in this scenario distribution of X is that of normal mixture for which no explicit formulae for its entropy exist (see, however, [34] for special cases of mixtures of summands having the same covariance matrix). Authors of [78] use approximations of MI (see (18) in [78]) to address this problem.

To summarise, we note that the section discusses possible approaches to test conditional independence and discover Markov Blanket of target Y when the underlying distribution is continuous. The first task can be performed by using some available estimators of CMI, discussed above, such as Kraskov et al. estimator [74] in conjunction with benchmark distribution based on resampling. The other possibility is a kernel method which reduces the problem to checking whether generalised conditional covariance is 0. For the parametric case when (X, Y) is Gaussian, the testing problem can be reduced to testing whether the partial correlation coefficient is 0. Moreover, in this case under faithfulness assumption on the distribution Markov Blanket can be recovered with theoretical guarantees.

10. Interaction Detection

Detection of existing interactions between features in influencing the outcome Y is an important task of data analysis; primary examples being gene–gene interactions in Genome-Wide Association Studies (GWAS) and gene–environment interactions. In this section, we show how information-theoretic measures introduced before can be applied to solve this problem. Moreover, we indicate that interaction detection problems may be viewed as a special case of feature selection. Various indices have been proposed to measure the strength of interactions, one of the most popular tools being interaction information discussed in Sections 2 and 4.1. There it was considered as a tool to define new selection criteria by truncating and possibly weighing summands in the Möbius expansion of CMI.

Here, II is adopted as a main tool in detection of interactions. Its most popular competitor is the value of Likelihood Ratio Test statistics for two nested logistic models: an additive model with no interactions and a saturated model taking all possible interactions into account. However, it has been shown in [68] that for logistic model when predictors are independent and at least one of interaction terms is non-zero then $II \neq 0$ but not vice versa. This shows that there are situations when interactions are present and are detected by II but will go undetected using logistic model approach. II is applied as the main tool to find interactions in AMBIENCE package [79] and BOOST package uses so called Kirkwood approximation discussed below, which is closely related to II [80]. Other competitors to LRT and II include Multifactor Dimensionality Reduction (MDR) [81] and Restricted Partitioning Method (RPM) [82].

Now, we will discuss two additional properties of Interaction Information which are useful in interaction analysis, focusing on its three-way variant applied to predict synergistic effect of two variables X_1 and X_2, say, in determining the target Y.

From the first equality in (5) it follows that when both variables are jointly independent from Y, i.e., $(X_1, X_2) \perp\!\!\!\perp Y$ then $II(X; X_2; Y) = 0$ as $I(X_1; X_2|Y) = I(X_1; X_2)$. Although the converse is not true, it can be shown in adversarially constructed examples only and, thus, testing $H_0 : II = 0$ is usually replaced by (more stringent) hypothesis $\tilde{H}_0 : (X_1, X_2) \perp\!\!\!\perp Y$.

The other property is insightful representation

$$II(X_1; X_2; Y) = KL(P_{X_1,X_2,Y}||\tilde{P}_K),$$

where $\tilde{P}_K$ is Kirkwood Superposition Approximation with masses assigned to points equal

$$\tilde{p}_K(x_1, x_2, y) = \frac{p(x_1, x_2)p(x_1, y)p(x_2, y)}{p(x_1)p(x_2)p(y)}$$

is positive but necessarily summing up to 1, mass function supported on values (x_1, x_2, y) of (X_1, X_2, Y). It can be shown that for η denoting the summary mass of $\tilde{P}_K$ it follows that when $\eta \leq 1$ and $II = 0$, then $P_{X_1,X_2,Y}$ is equal to its Kirkwood approximation [68].

Let us discuss two commonly used methods of testing that $II(X; X_2; Y) = 0$. The most popular one is based on Han's approximation [20] derived under a stringent assumption of overall independence, which results in considering as the benchmark distribution chi-squared distribution with $(I-1)(J-1)(K-1)$ degrees of freedom, when I, J, K are numbers of values of X_1, X_2, and Y, respectively. This, however, may lead to a large number of false signals when the pertaining test is employed as the overall independence is only a very special situation when II vanishes. It is shown in [69] that when (X_1, X_2) are independent of Y, similarly to other measures of conditional dependence discussed in Section 4.1, the approximate distribution is weighted chi-square distribution. Its complete description has been given for X_1 and X_2 having three values and Y being binary, which includes typical GWAS situation of two loci with two co-dominant alleles. The properties of the corresponding test have not been investigated yet.

The other method uses permuting Y values of the considered sample and in this way obtaining M random samples satisfying $\tilde{H}_0 : (X_1, X_2) \perp\!\!\!\perp Y$. Then, permutation p-value can be calculated as $(M+1)^{-1}(1 + \sum_{i=1}^{M} I(\hat{II}_i \geq \hat{II}))$ (compare (32)). One can also use chi-square distribution with the number of degrees of freedom as the benchmark distribution, analogously to [70]. The advantage of the latter approach is that the number of permuted samples M may be much smaller than in the case of permutation test.

The main challenge to detect significant interactions among potential predictors $F = \{1, \ldots, p\}$ is usually computational burden of the procedure. Indeed, any method discussed should in principle include interaction term for any pair of predictors $X_i, X_j, i, j \in F$. This requires enlarging set F by $|F| \times (|F| - 1)/2$ hot-encoded interactions, i.e., fitted model will contain $|F| \times (|F| + 1)/2$ nominal predictors. This is practically infeasible, e.g., for 100 Single Nucleotide Polymorphisms (SNP) being our predictors this would yield number of needed tests of order 10^6 at each step of selection procedure. The frequently adapted

approach is to screen the set of variables first and then apply more sophisticated procedure to the chosen variables only.

Consider two other possible solutions. The first one uses the premise that significant interactions may arise only among variables which themselves are significant. Thus any of the method described in Section 8 can be used for predictors in F and then two methods described above are applicable for all interactions of the chosen features using Bonferroni adjustment for multiple testing or Holm method [83]. The problematic aspect of this approach is exclusion of possible interactions between two features when only one of them has non-negligible main effect.

The other group of methods, for which BOOST proposed in [80] is a representative example, screens interactions by the modified version of LRT test which consists of replacing likelihood for a fitted additive model by likelihood for normalised Kirkwood approximation $\tilde{P}_K/\eta$ which can be very quickly computed. The pairs for which the value of LRT statistic modified in this way exceeds certain threshold. Then, an exact LRT test is performed for all remaining pairs. The weak side of this approach is a choice of a threshold τ in the first stage which needs to be chosen by a rule-of-thumb as properties of modified LRT test remain unknown.

11. Conclusions

This paper reviews main ideas concerning feature selection using information theoretic tools which revolve around detecting and measuring the strength of conditional independence. As estimation of CMI, which is a natural and powerful tool for this endeavour, encounters a problem when dimensionality of the task is large; a natural way is to look for its good substitutes. Several ways in which such substitutes are constructed are discussed and the current knowledge about their properties is presented. It is argued that selection criteria based on truncation of Möbius decomposition make quite far-reaching compromises on the dependence structures of feature sets, whereas properties of approximations based on variational approaches are yet to be established. Additionally, major Markov Blanket discovery algorithms are constructed under assumptions that conditional independence or its absence can be established without error, when, in reality, the intrinsic features of any practical CI test applied at each stage is its fallibility. Therefore, further efforts, both theoretical and experimental, are needed to understand the advantages, drawbacks, and interrelations between up-to-date developments. The paper presents some recently established tools which can be used for this purpose, such as asymptotic distributions of CMI-related criteria discussed in Section 6. They are used to construct tests of conditional independence with approximate control of probability of false signals. The problem of existence of interactions can be similarly approached using results in [69].

Another challenging task is to extend general approaches discussed here, such as construction of feature selection criteria by truncation of Möbius expansion or variational approach to multilabel classification when Y is a multivariate binary vector. Several criteria, such as $CIFE$ or JMI, have been generalised to this case already (cf. [84,85], respectively, see also [86] for an approach based on Han's inequality and [87] for the review), but the general approach, e.g., in the spirit of [24], is still missing. Another important challenge here seems construction of feature selection criteria which will efficiently take into account dependence between coordinates of the response.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Proof of Theorem 3.

Proof. (i) follows from the first equality in (16) after noting that the naive Bayes assumption implies in particular that: $X_S \perp\!\!\!\perp X|Y$ and, thus $I(X;X_S|Y) = 0$. Moreover we have

$$I(X;X_S) = H(X_S) - \sum_{i\in S} H(X_i|X) = H(X_S) - \sum_{i\in S} H(X_i) + \sum_{i\in S} I(X;X_i).$$

The conclusion follows as two first terms do not depend on X. (ii) follows directly from Möbius formula and in order to prove (iii) using $I(X;Y) = H(Y) - H(Y|X)$ we have

$$\begin{aligned} I(Y;X|X_S) &= I(X;Y) - I(X;X_S) + I(X;X_S|Y) \\ &= I(X;Y) - H(X_S) + H(X_S|X) + H(X_S|Y) - H(X_S|X,Y) \end{aligned} \tag{A1}$$

and, thus, up to terms which do not depend on X it equals in view of assumptions

$$I(X;Y) + H(X_S|X) - H(X_S|X,Y) = I(X;Y) + \sum_{i\in S}(H(X_i|X) - H(X_i|X,Y)) =$$
$$I(X;Y) + \sum_{i\in S} I(X_i;Y|X),$$

which differs from $CIFE(Y,X|X_S)$ by yet another term which does not depend on X.

(iv) We show now reasoning leading to Joint Mutual Information Criterion JMI (cf. [33,48]). Namely, we have for $i \in S$

$$I(X;X_S) = I(X;X_i) + I(X;X_{S\setminus\{i\}}|X_i).$$

Summing these equalities over all $i \in S$ and dividing by $|S|$ we obtain

$$I(X;X_S) = \frac{1}{|S|}\sum_{i\in S} I(X;X_i) + \frac{1}{|S|}\sum_{i\in S} I(X;X_{S\setminus\{i\}}|X_i)$$

and analogously

$$I(X;X_S|Y) = \frac{1}{|S|}\sum_{i\in S} I(X;X_i|Y) + \frac{1}{|S|}\sum_{i\in S} I(X;X_{S\setminus\{i\}}|X_i,Y).$$

Subtracting two last equations and using definition of II we obtain

$$I(Y;X|X_S) = I(X;Y) + \frac{1}{|S|}\sum_{i\in S} II(X;X_i;Y) + \frac{1}{|S|}\sum_{i\in S} II(X;X_{S\setminus\{i\}};Y|X_i).$$

Moreover it follows from definition of II that when X is independent from $X_{S\setminus\{i\}}$ given X_i and these quantities are independent given X_i and Y the last sum is 0 and we obtain definition of JMI. □

References

1. Hinton, G.; Roweis, S. Stochastic neighbor embedding. In Proceedings of the Neural Information Processing Systems NIPS2002, Vancouver, BC, Canada, 9–14 December 2002.
2. Faivishevsky, L.; Goldberger, J. A nonparametric information theoretic clustering algorithm. In Proceedings of the ICML, Haifa, Israel, 21–24 June 2010.
3. Izenman, A. *Modern Multivariate Statistical Techniques*; Springer: New York, NY, USA, 2008.
4. Bishop, C. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
5. Dębowski, L. *Information Theory Meets Power Laws*; Wiley: Hoboken, NJ, USA, 2020.
6. Hastie, R.; Friedman, J.; Tibshirani, R. *Elements of Statistical Learning*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2009.
7. Hinton, G.; Salakhutdinov, R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507.
8. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

9. Draminski, M.; Koronacki, J. RMCFS: An R package for Monte Carlo feature selection and interdependency discovery. *J. Stat. Softw.* **2018**, *85*, 1–28.
10. Kursa, M.; Rudnicki, W. Feature selection with Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–11.
11. Guyon, I.; Eliseef, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
12. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv.* **2018**, *50*, 1–45.
13. Macedo, F.; Rosário de Oliveira, M.; Pacheco, A.; Valadas, R. Theoretical foundations of forward feature selection methods based on mutual information. *Neurocomputing* **2019**, *325*, 67–89.
14. Yu, K.; Liu, L.; Li, J. A Unified view of causal and non-causal feature selection. *ACM Trans. Knowl. Discov. Data* **2021**, *15*, 1–46.
15. Cover, T.M.; Thomas, J.A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*; Wiley-Interscience: Hoboken, NJ, USA, 2006.
16. Yeung, R.W. *A First Course in Information Theory*; Kluwer: New York, NY, USA, 2002.
17. Konishi, S.; Kitagawa, G. *Information Criteria and Statistical Modeling*; Springer: New York, NY, USA, 2009.
18. McGill, W.J. Multivariate information transmission. *Psychometrika* **1954**, *19*, 97–116.
19. Ting, H.K. On the amount of information. *Theory Probab. Appl.* **1960**, *7*, 439–447.
20. Han, T.S. Multiple mutual informations and multiple interactions in frequency data. *Inf. Control* **1980**, *46*, 26–45.
21. Fano, R. *Transmission of Information*; MIT Press: Cambridge, MA, USA, 1961.
22. Kolchinsky, A.; Tracey, B.; Wolpert, D. Nonlinear information bottleneck. *Entropy* **2019**, *21*, 1181.
23. Meyer, P.; Schretter, C.; Bontempi, G. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE J. Sel. Top. Signal Process.* **2008**, *2*, 261–274.
24. Brown, G.; Pocock, A.; Zhao, M.J.; Luján, M. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* **2012**, *13*, 27–66.
25. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks for Plausible Inference*; Morgan Kaufmann: San Francisco, CA, USA, 1988.
26. Candès, E.; Fan, Y.; Janson, Y.; Lv, J. Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. (Stat. Methodol.)* **2016**, *80*, 551–577.
27. Margaritis, D. Towards provably correct feature selection in arbitrary domains. In Proceedings of the 22th International Conference on Neural Information Processing Systems (NIPS'09), Vancouver, BC, Canada, 7–10 December 2009.
28. Nilsson, R.; M. Peña, J.; Björkegren, J.; Tegnér, J. Consistent Feature Selection for Pattern Recognition in Polynomial Time. *J. Mach. Learn. Res.* **2007**, *8*, 589–612.
29. Kubkowski, M.; Mielniczuk, J.; Teisseyre, P. How to gain on power: Novel conditional independence tests based on short expansion of conditional mutual information. *J. Mach. Learn. Res.* **2021**, *22*, 1–57.
30. Lewis, D. Feature selection and feature extraction for text categorization. In Proceedings of the Workshop on Speech and Natural Language, Harriman, NY, USA, 23–26 February 1992; pp. 212–217.
31. Lin, D.; Tang, X. Conditional Infomax Learning: An integrated framework for feature extraction and fusion. In Proceedings of the 9th European Conference on Computer Vision (ECCV'06)—Volume Part I, Graz, Austria, 7–13 May 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 68–82.
32. Pawluk, M.; Teisseyre, P.; Mielniczuk, J. Information-theoretic feature selection using high-order interactions. In Proceedings of the Machine Learning, Optimization, and Data Science, Volterra, Italy, 13–16 September 2018; Springer International Publishing: Cham, Switzerland, 2019; pp. 51–63.
33. Yang, H.H.; Moody, J. Data visualization and feature selection: New algorithms for nongaussian data. *Adv. Neural Inf. Process. Syst.* **1999**, *12*, 687–693.
34. Łazęcka, M. Mielniczuk, J. Analysis of information-based nonparametric variable selection criteria. *Entropy* **2020**, *22*, 974.
35. Battiti, R. Using mutual information for selecting features in supervised net learning. *IEEE Trans. Neural Netw.* **1994**, *5*, 537–550.
36. Peng, H.; Long, F.; Ding, C. Feature selection based on mutula information: Criteria of max-dependency,max-relevance, and min-dependency. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238.
37. Estevez, P.; Tesmer, M.; Perez, C.; Zurada, J. Normalized mutual information feature selection. *IEEE Trans. Neural Netw.* **2009**, *20*, 189–201.
38. Sechidis, K.; Azzimonti, L.; Pocock, A.; Corani, G.; Weatherall, J.; Brown, G. Efficient feature selection using shrinkage estimators. *Mach. Learn.* **2019**, *108*, 1261–1286.
39. Fleuret, F. Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* **2004**, *5*, 1531–1555.
40. Shishkin, A.; Bezzubtseva, A.; Drutsa, A.; Shishkov, I.; Gladkikh, E.; Gusev, G.; Serdyukov, P. Efficient high-order interaction-aware Feature selection based on Conditional Mutual Information. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16), Barcelona, Spain, 5–10 December 2016; pp. 4644–4652.
41. Donsker, M.; Varadhan, S. Asymptotic evaluation of certain Markov process expectations for large time. IV. *Commun. Pure Appl. Math.* **1983**, *36*, 183–212.
42. Belghazi, M.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, D. Mutual information neural estimation. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 531–540.

43. Poole, B.; Ozair, S.; Oord, A.; Alemi, A.; Tucker, G. On variational bounds of mutual information. In Proceedings of the ICML Proceedings, PMLR 97, Long Beach, CA, USA, 9–15 June 2019; pp. 1–12.
44. Molavipour, S.; Bassi, G.; Skoglund, M. Conditional mutual information neural estimator. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 5025–5029.
45. Nguyen, X.; Wainwright, M.; Jordan, M. Estimating divergence functionals and the likelihood ratio by convex risk minimisation. *IEEE Trans. Inf. Theory* **2010**, *56*, 5847–5861.
46. Gao, S.; Ver Steer, G.; Galstyan, A. Variational information Maximisation for Feature Selection. In Proceedings of the 30th Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 487–495.
47. Bach, F.; Jordan, M. Beyond independent components: Trees and clusters. *J. Mach. Learn. Res.* **2003**, *7*, 1205–1233.
48. Vergara, J.; Estévez, P. A review of feature selection methods based on mutual information. *Neural. Comput. Appl.* **2014**, *24*, 175–186.
49. Vinh, N.; Zhou, S.; Chan, J.; Bailey, J. Can high-order dependencies improve mutual information based feature selection? *Pattern Recognit.* **2016**, *53*, 45–58.
50. Łazęcka, M.; Mielniczuk, J. Squared error-based shrinkage estimators of discrete probabilities and their application to variable selection. *Stat. Pap.* **2022**, 1261–1286. https://doi.org/10.1007/s00362-022-01308-w.
51. Kullback, S. *Information Theory and Statistics*; Peter Smith: Cloucester, MA, USA, 1978.
52. Shao, I. *Mathematical Statistics*; Springer: Berlin/Heidelberg, Germany, 2003.
53. Agresti, A. *Categorical Data Analysis*; Wiley: Hoboken, NJ, USA, 2002.
54. Łazęcka, M. Mielniczuk, J. Multiple testing of conditional independence using information theoretic-approach. In Proceedings of the Modelling Decisions for Artificial Intelligence'2021, LNAI 12898, Umeå, Sweden, 27–30 September 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 1–12.
55. Kubkowski, M.; Łazęcka, M.; Mielniczuk, J. Distributions of a general reduced-order dependence measure and conditional independence testing. In Proceedings of the International Conference on Computational Science ICCS'20, Amsterdam, The Netherlands, 3–5 June 2020; Springer International Publishing: Cham, Switzerland, 2020, pp. 692–706.
56. Zhang, J.T. Approximate and Asymptotic distributions of chi-squared type mixtures with applications. *J. Am. Stat. Assoc.* **2005**, *100*, 273–285.
57. Barber, R.; Candès, E. Controlling the false discovery rate via knockoffs. *Ann. Stat.* **2015**, *43*, 2055–2085.
58. Berrett, T.; Wang, Y.; Barber, R.; Samworth, R. The conditional permutation test for independence while controlling for confounders. *J. R. Stat. Soc. Ser. (Stat. Methodol.)* **2020**, *82*, 175–197.
59. Sen, R.; Suresh, A.; Shanmugam, K.; Dimakis, A.; Shakkottai, S. Model-powered conditional independence test. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 2951–2961.
60. Koller, D.; Sahami, M. Toward optimal feature selection. In Proceedings of the ICML-1995, Tahoe City, CA, USA, 9–12 July 1995; pp. 284–292.
61. Aliferis, C.; Tsamardinos, I.; Statnikov, A. HITON: A novel Markov Blanket algorithm for optimal variable selection. In Proceedings of the AMIA Annual Symposium Proceedings, Washington, DC, USA, 8–12 November 2003; pp. 21–25.
62. Tsamardinos, I.; Aliferis, C.; Statnikov, A. Time and sample efficient discovery of Markov Blankets and direct causal relations. In Proceedings of the 9th ACM SIGD Conference on KDDM, Washington, DC, USA, 24–27 August 2003; pp. 673–678.
63. Fu, S.; Desormais, M. Fast Markov Blanket discovery algorithm via local learning within single pass. In Proceedings of the CSCSI Conference, Las Vegas, USA, 14–16 December 2017; pp. 96–107.
64. Gao, T.; Qiang, J. Efficient Markov blanket discovery and its application. *IEEE Trans. Cybern.* **2017**, *47*, 1169–1179.
65. Margaritis, D.; Thrun, S. Bayesian network induction via local neighborhoods. In Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS'99), Denver, CO, USA, 29 November–4 December 1999; pp. 505–511.
66. Tsamardinos, I.; Aliferis, C.F.; Statnikov, A.R. Algorithms for large scale Markov blanket discovery. In Proceedings of the FLAIRS Conference, St. Augustine, FL, USA, 12–14 May 2003; pp. 376–381.
67. Bühlmann, P.; van de Geer, S. *Statistics for High-Dimensional Data*; Springer: Berlin/Heidelberg, Germany, 2011.
68. Mielniczuk, J.; Teisseyre, P. A deeper look at two concepts of measuring gene–gene interactions: logistic regression and interaction information revisited. *Genet. Epidemiol.* **2018**, *42*, 187–200.
69. Kubkowski, M.; Mielniczuk, J. Asymptotic distributions of interaction information. *Methodol. Comput. Appl. Probab.* **2021**, *23*, 291–315.
70. Tsamardinos, I.; Borboudakis, G. Permutation testing improves on Bayesian network learning. In Proceedings of the ECML PKDD 2010, Barcelona, Spain, 20–24 September 2010; pp. 322–337.
71. Shah, R.; Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *Ann. Stat.* **2018**, *48*, 1514–1538.
72. Kozachenko, L.; Leonenko, N. Sample estimate of entropy of a random vector. *Probl. Inf. Transm.* **1987**, *23*, 95.
73. Berrett, T.; Samworth, R. Nonparametric independence testing via mutual information. *Biometrika* **2019**, *106*, 547–566.
74. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138.
75. Daudin, J. Partial association measures and application to anqualitative regression. *Biometrika* **1980**, *67*, 581–590.

76. Liu, M.; Katsevich, E.; Janson, L.; Ramdas, A. Fast and powerful conditional randomization testing via distillation. *Biometrika* **2022**, *109*, 277–293.
77. Sprites, P.; Glymour, C.; Scheines, R. *Causation, Prediction and Search*; MIT Press: Cambridge, MA, USA, 2000.
78. Lefakis, L.; Fleuret, F. Jointly informative feature selection made tractable by gaussian modeling. *J. Mach. Learn. Res.* **2016**, *17*, 1–39.
79. Chanda, P.; Sucheston, L.; Zhang, A.; Brazeau, D.; Freudenheim, J.; Ambrosone, C.; Ramanathan, M. AMBIENCE: A novel Approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. *Genetics* **2008**, *180*, 1191–1210.
80. Wan, X.; Yang, C.; Yang, Q.; Xue, H.; Fan, X.; Tang, N.; Yu, W. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.* **2010**, *87*, 325–340.
81. Ritchie, M.D.; Hahn, L.W.; Roodi, N.; Bailey, L.R.; Dupont, W.D.; Parl, F.F.; Moore, J.H. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **2001**, *69*, 138–147.
82. Culverhouse, R. The use of the restricted partition method with case-control data. *Hum. Hered.* **2007**, *93–100*, 138–147.
83. Dudoit, S.; Laan, M. *Multiple Testing Procedures with Application to Genomics*; Springer: New York, NY, USA, 2008.
84. Lee, J.; Kim, D.W. Feature selection for multi-label classification using multivariate mutual information. *Pattern Recognit. Lett.* **2013**, *34*, 349–357.
85. Sechidis, K.; Nikolaou, N.; Brown, G. Information theoretic feature selection in multi-label data through composite likelihood. In Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, Joensuu, Finland, 20–22 August 2014; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8621, pp. 143–152.
86. Wangduk Seo, Dae-Won Kim, J.L. Generalized information-theoretic criterion for multi-label feature selection. *IEEE Access* **2019**, *7*, 122854–122863.
87. Kashef, S.; Nezamabadi-pour, H.; Nikpour, B. Multilabel feature selection: A comprehensive review and guiding experiments. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, 1–29.