entropy

MDPI

Article

Analysis of Information-Based Nonparametric Variable Selection Criteria

Małgorzata Łazęcka [1,2] and Jan Mielniczuk [1,2,*]

[1] Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warsaw, Poland; m.lazecka@ipipan.waw.pl

[2] Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland

* Correspondence: miel@ipipan.waw.pl

Received: 6 August 2020; Accepted: 28 August 2020; Published: 31 August 2020

Abstract: We consider a nonparametric Generative Tree Model and discuss a problem of selecting active predictors for the response in such scenario. We investigated two popular information-based selection criteria: Conditional Infomax Feature Extraction (CIFE) and Joint Mutual information (JMI), which are both derived as approximations of Conditional Mutual Information (CMI) criterion. We show that both criteria CIFE and JMI may exhibit different behavior from CMI, resulting in different orders in which predictors are chosen in variable selection process. Explicit formulae for CMI and its two approximations in the generative tree model are obtained. As a byproduct, we establish expressions for an entropy of a multivariate gaussian mixture and its mutual information with mixing distribution.

Keywords: conditional mutual information; CMI; information measures; nonparametric variable selection criteria; gaussian mixture; conditional infomax feature extraction; CIFE; joint mutual information criterion; JMI; generative tree model; Markov blanket

1. Introduction

In the paper, we consider theoretical properties of Conditional Mutual Information (CMI) and its approximations in a certain dependence model called Generative Tree Model (GTM). CMI and its modifications are used in many problems of machine learning including feature selection, variable importance ranking, causal discovery, and structure learning of dependence networks (see, e.g., Reference [1,2]). They are the cornerstone of nonparametric methods to solve such problems meaning that no parametric assumptions on dependence structure are imposed. However, formal properties of these criteria remain largely unknown. This is mainly due to two problems: firstly, theoretical values of CMI and related quantities are hard to calculate explicitly, especially when the conditioning set has a large dimension. Moreover, there are only a few established facts about behavior of their sample counterparts. Such a situation, however, has important consequences. In particular, a relevant question whether certain information based criteria, such as Conditional Infomax Feature Extraction (CIFE) and Joint Mutual Information (JMI), obtained as approximations of CMI, e.g., by truncation of its Möbius expansion are approximations in analytic sense (i.e., whether the difference of both quantities is negligible) remains unanswered. In the paper, we try to fill this gap. The considered GTM is a model for which marginal distributions of predictors are mixtures of gaussians. Exact values of CMI, as well as of those of CIFE and JMI, are calculated for this model, which makes studying their behavior when parameters of the model and number of predictors change feasible. In particular, it is shown that CIFE and JMI exhibit different behavior than CMI and also they may significantly differ between themselves. In particular, we show, that depending on the value of model parameters, each of considered criteria

JMI and CIFE can incorporate inactive variables before active ones into a set of chosen predictors. This, of course, does not mean that important performance criteria, such as False Detection Rate (FDR), cannot be controlled for CIFE and JMI but should serve as a cautionary note that their similarity to CMI, despite their derivation, is not necessarily ensured. As a byproduct, we establish expressions for an entropy of a multivariate gaussian mixture and its mutual information with mixing distribution, which are of independent interest.

We stress that our approach is intrinsically nonparametric and focuses on using nonparametric measures of conditional dependence for feature selection. By studying their theoretical behavior for this task we also learn an average behavior of their empirical counterparts for large sample sizes. Generative Tree Model appears, e.g., in Reference [3], a non-parametric tree structured model is also considered, e.g., in Reference [4,5]. Together with autoregressive model, it is one of the two most common types of generative models. Besides its easily explainable dependence structure, distributions of predictors in the considered model are mixed gaussians, and this facilitates calculation of explicit form of information-based selection criteria.

The paper is structured as follows. Section 2 contains information-theoretic preliminaries, some necessary facts on information based feature-selection and derivation of CIFE and JMI criteria as approximations of CMI. Section 3 contains derivation of entropy and mutual information for gaussian mixtures. In Section 4, behavior of CMI, CIFE, and JMI is studied in GTM. Section 5 concludes.

2. Preliminaries

We denote by $p(x)$, $x \in \mathbb{R}^d$ a probability density function corresponding to continuous variable X on $\mathbb{R}^d$. Joint density of X and variable Y will be denoted by $p(x, y)$. In the following, Y will denote discrete random response to be predicted using multivariate vector X.

Below, we discuss some information-theoretic preliminaries, which leads, at the end of Section 2.1, to Möbius decomposition of mutual information. This is used in Section 2.2 to construct CIFE approximation of CMI. In addition, properties of Mutual Information discussed in Section 2.1 are used in Section 2.2 to justify JMI criterion.

2.1. Information-Theoretic Measures of Dependence

The (differential) entropy for continuous random variable X is defined as

$$H(X) = -\int_{\mathbb{R}^d} p(x) \log p(x)\, dx \tag{1}$$

and quantifies the uncertainty of observing random values of X. Note that the definition above is valid regardless the dimensionality d of the range of X. For discrete X, we replace the integral in (1) by the sum and density $p(x)$ by probability mass function. In the following, we will frequently consider subvectors of $X = (X_1, \ldots, X_p)$, which is a vector of all potential predictors of discrete response Y. The conditional entropy of X given discrete Y is written as

$$H(X|Y) = \sum_{y \in \mathcal{Y}} p(y) H(X|Y = y). \tag{2}$$

When Z is continuous, the conditional entropy $H(X|Z)$ is defined as $\mathbb{E}_Z H(X|Z = z)$, i.e.,

$$H(X|Z) = -\int p(z) \int \frac{p(x,z)}{p(z)} \log\left(\frac{p(x,z)}{p(z)}\right) dx dz = -\int p(x,z) \log\left(\frac{p(x,z)}{p(z)}\right) dx dz, \tag{3}$$

where $p(x, z)$ and $p(z)$ denote joint density of (X, Z) and density of Z, respectively. The mutual information (MI) between X and Y is

$$I(X, Y) = H(X) - H(X|Y) = H(X) - H(Y|X). \tag{4}$$

This can be interpreted as the amount of uncertainty in X (Y) which is removed when Y (respectively, X) is known, which is consistent with the intuitive meaning of mutual information as the amount of information that one variable provides about another. It determines how similar the joint distribution is to the product of marginal distributions when Kullback-Leibler divergence is used as similarity measure (cf. Reference [6], Equation (8.49)). Thus, $I(X,Y)$ may be viewed as nonparametric measure of dependence. Note that, as $I(X,Y)$ is symmetric, it only shows the strength of dependence but not its direction. In contrast to correlation coefficient MI is able to discover non-linear relationships as it equals zero if and only if X and Y are independent. It is easily seen that $I(X,Y) = H(X) + H(Y) - H(X,Y)$. A natural extension of MI is conditional mutual information (CMI) defined as

$$I(X,Y|Z) = H(X|Z) - H(X|Y,Z) = \int p(z) \int p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \, dxdydz, \tag{5}$$

which measures the conditional dependence between X and Y given Z. When Z is a discrete random variable, the first integral is replaced by a sum. Note that the conditional mutual information is mutual information of X and Y given $Z = z$ averaged over values z of Z, and it equals zero if and only if X and Y are conditionally independent given Z. Important property of MI is a chain rule which connects $I((X_1, X_2), Y)$ with $I(X_1, Y)$:

$$I((X_1, X_2), Y) = I(X_1, Y) + I(X_2, Y|X_1). \tag{6}$$

For more properties of the basic measures described above, we refer to Reference [6,7]. We define now interaction information II ([8]), which is a useful tool for decomposing mutual information between multivariate random variable X_S and Y (see formula (13) below). The 3-way interaction information is defined as

$$II(X_1, X_2, Y) = I((X_1, X_2), Y) - I(X_1, Y) - I(X_2, Y). \tag{7}$$

This is frequently interpreted as the part of $I((X_1, X_2), Y)$, which remains after subtraction of individual informations between Y and X_1 and Y and X_2. The definition indicates in particular that $II(X_1, X_2, Y)$ is symmetric. Note that it follows from (6) that

$$II(X_1, X_2, Y) = I(X_1, Y|X_2) - I(X_1, Y) = I(X_2, Y|X_1) - I(X_2, Y), \tag{8}$$

which is consistent with the intuitive meaning of existence of interaction as a situation in which the effect of one variable on the class variable Y depends on the value of another variable. By expanding all mutual informations on RHS of (7), we obtain

$$II(X_1, X_2, Y) = -H(X_1) - H(X_2) - H(Y) + H(X_1, Y) + H(X_2, Y) + H(X_1, X_2) - H(X_1, X_2, Y). \tag{9}$$

The 3-way II can be extended to the general case of p variables. The p-way interaction information [9,10] is

$$II(X_1, \ldots, X_p) = -\sum_{T \subseteq \{1,\ldots,p\}} (-1)^{p-|T|} H(X_T). \tag{10}$$

For $p = 2$, (10) reduces to mutual information, whereas, for $p = 3$, it reduces to (9).

We consider two useful properties of introduced measures. We first start with 3-way information interaction, and we note that it inherits chain-rule property from MI, namely

$$II(X_1, (X_2, X_3), Y) = II(X_1, X_3, Y) + II(X_1, X_2, Y|X_3), \tag{11}$$

where $I(X_1, X_2, Y|X_3)$ is defined analogously to (7) by replacing mutual informations on RHS by conditional mutual informations given X_3. This is easily proved by writing, in the view of (6):

$$II(X_1, (X_2, X_3), Y) = I(X_1, (X_2, X_3)|Y) - I(X_1, (X_2, X_3)) =$$

$$I(X_1, X_3|Y) + I(X_1, X_2|Y, X_3) - [I(X_1, X_3) + I(X_1, X_2|X_3)] \tag{12}$$

and using (8) in the above equalities. Namely, joining the first and the third expression together (and the second and the fourth, as well), we obtain that RHS equals $II(X_1, X_3, Y) + II(X_1, X_2, Y|X_3)$.

We also state Möbius representation of mutual information which plays an important role in the following development. For $S \subseteq \{1, 2, \ldots, p\}$, let X_S be a random vector coordinates of which have indices in S. Möbius representation [10–12] states that $I(X_S, Y)$ can be recovered from interaction informations

$$I(X_S, Y) = \sum_{k=1}^{|S|} \sum_{\{t_1,\ldots,t_k\} \subseteq S} II(X_{t_1}, \ldots, X_{t_k}, Y), \tag{13}$$

where $|S|$ denotes number of elements of set S.

2.2. Information-Based Feature Selection

We consider discrete class variable Y and p features $X_1, \ldots, X_p$. We do not impose any assumptions on dependence between Y and $X_1, \ldots, X_p$, i.e., we view its distributional structure in a nonparametric way. Let X_S denote a subset of features, indexed by set $S \subseteq \{1, \ldots, p\}$. As $I(X_S, Y)$ does not decrease when S is replaced by its superset $S' \supseteq S$, the problem of finding $\arg\max_S I(X_S, Y)$ has a trivial solution $full = \{1, 2, \ldots, p\}$. Thus, one usually tries to optimize mutual information between X_S and Y under some constraints on the size $|S|$ of S. The most intuitive approach is an analogue of k-best subset selection in regression which tries to identify a feature subset of a fixed size $1 \leq k \leq p$ that maximizes the joint mutual information with a class variable Y. However, this is infeasible for large k because the search space grows exponentially with the number of features. As a result, various greedy algorithms have been developed including forward selection, backward elimination and genetic algorithms. They are based on observation that

$$\arg\max_{j \in S^c} [I(X_{S \cup \{j\}}, Y) - I(X_S, Y)] = \arg\max_{j \in S^c} I(X_j, Y|X_S), \tag{14}$$

where $S^c = \{1, \ldots, p\} \setminus S$ is a complement of S. The equality in (14) follows from (6). In each step, the most promising candidate is added. In the case of ties in (14), the variable satisfying it with the smallest index is chosen.

2.3. Approximations of CMI: CIFE and JMI Criteria

Observe that it follows from (13)

$$I(X_{S \cup \{j\}}, Y) - I(X_S, Y) = I(X_j, Y|X_S) = \sum_{k=0}^{|S|} \sum_{\{t_1,\ldots,t_k\} \subseteq S} II(X_{t_1}, \ldots, X_{t_k}, X_j, Y). \tag{15}$$

Direct application of the above formula to find the maximizer in (14) is infeasible as estimation of a specific information interaction of order k requires $O(C^k)$ observations. The above formula allows us, however, to obtain various natural approximations of CMI. The first order approximation does not take interactions between features into account and that is why the second order approximation obtained by taking first two terms in (15) is usually considered. The corresponding score for candidate feature X_j is

$$CIFE(X_j, Y|X_S) = I(X_j, Y) + \sum_{i \in S} II(X_i, X_j, Y) = I(X_j, Y) + \sum_{i \in S} \left[I(X_i, X_j|Y) - I(X_i, X_j) \right]. \tag{16}$$

The acronym CIFE stand for Conditional Infomax Feature Extraction, and the measure has been introduced in Reference [13]. Observe that if interactions of order 3 and higher between predictors are 0, i.e., $II(X_{t_1}, \ldots, X_{t_k}, X_j, Y) = 0$ for $k \geq 2$ and then CIFE coincides with CMI. In Reference [2], it

is shown that CMI also coincides with CIFE if certain dependence assumptions on vector (X, Y) are satisfied. In view of the discussion above, CIFE can be viewed as a natural approximation to CMI. Observe that, in (16), we take into account not only relevance of the candidate feature, but also the possible interactions between the already selected features and the candidate feature. The empirical evaluation indicates that (16) is among the most successful MI-based methods; see Reference [2] for an extensive comparison of several MI-based feature selection approaches. We mention in this context, Reference [14], in which stopping rules for CIFE-based methods are considered.

Some additional assumptions lead to other score functions. We show now reasoning leading to Joint Mutual Information Criterion JMI (cf. Reference [12], on which the derivation below is based). Namely, if we define $S = \{j_1, \ldots, j_{|S|}\}$, we have for $i \in S$

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

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

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

and analogously

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

Subtracting the two last equations and using (8), we obtain

$$I(X_j, Y|X_S) = I(X_j, Y) + \frac{1}{|S|}\sum_{i\in S} II(X_j, X_i, Y) + \frac{1}{|S|}\sum_{i\in S} II(X_j, X_{S\setminus\{i\}}, Y|X_i). \tag{17}$$

Moreover, it follows from (8) that when X_j is independent of $X_{S\setminus\{i\}}$ given X_i and these quantities are independent given X_i and Y the last sum is 0 and we obtain equality

$$JMI(X_j, Y|X_S) = I(X_j, Y) + \frac{1}{|S|}\sum_{i\in S} II(X_j, X_i, Y) = I(X_j, Y) + \frac{1}{|S|}\sum_{i\in S}\left[I(X_j, X_i|Y) - I(X_j, X_i)\right]. \tag{18}$$

This is Joint Mutual Information Criterion (JMI) introduced in Reference [15]. Note that (18) together with (8) imply another useful representation

$$JMI(X_j, Y|X_S) = I(X_j, Y) + \frac{1}{|S|}\sum_{i\in S}\left[I(X_j, Y|X_i) - I(X_j, Y)\right] = \frac{1}{|S|}\sum_{i\in S} I(X_j, Y|X_i). \tag{19}$$

JMI can be viewed as an approximation of CMI when independence assumptions on which the above derivation was based are satisfied only approximately. Observe that $JMI(X_j, Y|X_S)$ differs from $CIFE(X_j, Y|X_S)$ in that the influence of the sum of interaction informations $II(X_j, X_i, Y)$ is down weighted by factor $|S|^{-1}$ instead of 1. This is sometimes interpreted as coping with 'redundancy over-scaled' problem (cf. Reference [2]). When the terms $I(X_j, X_i|Y)$ are omitted from the sum above then minimal redundancy maximal relevance (mRMR) criterion is obtained [16]. We note that approximations of CMI, such as CIFE or JMI, can be used in place of CMI in (14). As the derivation in both cases is quite intuitive, it is natural to ask how the approximations compare when used for selection. This is the primary aim of the present paper. Theoretical behavior of such methods will be investigated in the following sections. Note that we do not consider empirical counterparts of the above selection rules and investigate how they would behave provided their values have been known exactly.

3. Auxiliary Results: Information Measures for Gaussian Mixtures

In the following section, we will prove some results on information-theoretic properties of gaussian mixtures which are necessary to analyze the behavior of CMI, CIFE, and JMI in Generative Tree Model defined below.

In the next section, we will consider a gaussian Generative Tree Model, in which the main components have marginal distributions being mixtures of normal distributions. Namely, if Y has Bernoulli distribution $Y \sim \text{Bern}\,(1/2)$ (i.e., it admits values 0 and 1 with probability $1/2$) and distribution of X is defined as $\mathcal{N}\,(\mu Y, \Sigma)$, then X is a mixture of two normal distributions: $\mathcal{N}\,(0, \Sigma)$ and $\mathcal{N}\,(\mu, \Sigma)$ with equal weights. Thus, in this section, we state auxiliary results on entropy of such random variable and its mutual information with its mixing distribution. The result for entropy of multivariate gaussian mixture, to the best of our knowledge, is new; for univariate case, it was derived in Reference [17]. Bounds and approximations of the entropy of a gaussian mixture are used, e.g., in signal processing; see, e.g., Reference [18,19]. Consider d-dimensional gaussian mixture X defined as

$$X \sim \frac{1}{2}\mathcal{N}\,(0, I_d) + \frac{1}{2}\mathcal{N}\,(\mu, I_d)\,, \tag{20}$$

where '$\sim$' signifies 'distributed as'.

Theorem 1. *Differential entropy of X in (20) equals*

$$H(X) = h(\|\mu\|) + \frac{d-1}{2}\log(2\pi e),$$

where $h(a)$ is the differential entropy of one-dimensional gaussian mixture $2^{-1}\{\mathcal{N}(0,1) + \mathcal{N}(0,a)\}$ for $a > 0$.

$$h(a) = -\int\limits_{\mathbb{R}} \frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}} + e^{-\frac{(x-a)^2}{2}}\right)\log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}} + e^{-\frac{(x-a)^2}{2}}\right)\right)dx. \tag{21}$$

Proof. In order to avoid burdensome notation, we prove the theorem for $d = 2$ only. By the definition of differential entropy, we have

$$H(X) = -\int\limits_{\mathbb{R}^2} \frac{1}{2}\left(f_0(x_1,x_2) + f_\mu(x_1,x_2)\right)\log\left(\frac{1}{2}(f_0(x_1,x_2) + f_\mu(x_1,x_2))\right)dx_1dx_2,$$

where X is defined in (20) for $d = 2$, and f_μ denotes the density of normal distribution with a mean μ and a covariance matrix I_2.

We calculate the integral above changing the variables according to the following rotation

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \frac{\mu_1}{\|\mu\|} & -\frac{\mu_2}{\|\mu\|} \\ \frac{\mu_2}{\|\mu\|} & \frac{\mu_1}{\|\mu\|} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Transformed densities f_0 and f_μ are equal

$$f_0(y_1,y_2) = \frac{1}{2\pi}\exp\left(-\frac{y_1^2 + y_2^2}{2}\right)$$

and

$$f_\mu(y_1,y_2) = \frac{1}{2\pi}\exp\left(-\frac{(y_1 - \|\mu\|)^2 + y_2^2}{2}\right).$$

Applying above transformation, we can decompose $H(X)$ into sum of two integrals as follows:

$$H(X) = \int_{\mathbb{R}} \frac{1}{2\sqrt{2\pi}} \left(e^{-\frac{1}{2}y_1^2} + e^{-\frac{1}{2}(y_1 - \|\mu\|)^2} \right) \log \left(\frac{1}{2\sqrt{2\pi}} \left(e^{-\frac{1}{2}y_1^2} + e^{-\frac{1}{2}(y_1 - \|\mu\|)^2} \right) \right) dy_1$$
$$+ \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_2^2} \log \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_2^2} \right) dy_2 = h(\|\mu\|) + \frac{1}{2}\log(2\pi e),$$

where in the last equality the value $H(Z) = \log(2\pi e)/2$ for $N(0,1)$ variable Z is used. This ends the proof. □

The result above is now generalized to the case of arbitrary covariance matrix Σ. The general case will follow from Theorem 1 and the scaling property of differential entropy under linear transformations.

Theorem 2. *Differential entropy of*

$$X \sim \frac{1}{2}\mathcal{N}(0, \Sigma) + \frac{1}{2}\mathcal{N}(\mu, \Sigma)$$

equals

$$H(X) = h\left(\left\| \Sigma^{-1/2}\mu \right\| \right) + \frac{d-1}{2}\log(2\pi e) + \frac{1}{2}\log\left(\det \Sigma\right).$$

Proof. We apply Theorem 1 to multivariate random variable $Y = \Sigma^{-\frac{1}{2}}X$. We obtain

$$H(Y) = h\left(\left\| \Sigma^{-1/2}\mu \right\| \right) + \frac{d-1}{2}\log(2\pi e).$$

Using the scaling property of differential entropy [6], we have

$$H(X) = H(Y) + \frac{1}{2}\log(\det \Sigma),$$

which completes the proof. □

Similarly, we obtain the formula for mutual information of gaussian mixture and its mixing distribution. We use shorthand $X|Y = y$ to denote random variable defined as having distribution coinciding with conditional distribution $P(X|Y = y)$.

Theorem 3. *Mutual information of X and Y where $Y \sim$ Bern $(1/2)$ and $X|Y = y \sim \mathcal{N}(y\mu, \Sigma)$ equals*

$$I(X, Y) = h\left(\left\| \Sigma^{-1/2}\mu \right\| \right) - \frac{1}{2}\log(2\pi e). \tag{22}$$

Proof. We will use here the fact that the entropy of multidimensional normal distribution $Z \sim \mathcal{N}(\mu_Z, \Sigma)$ equals (cf. Reference [6], Theorem 8.4.1)

$$H(Z) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log(\det \Sigma).$$

Therefore, we have

$$I(X, Y) = H(X) - H(X|Y) = h\left(\left\| \Sigma^{-1/2}\mu \right\| \right) - \frac{1}{2}\log(2\pi e), \tag{23}$$

as

$$H(X|Y) = \frac{1}{2}H(X|Y = 0) + \frac{1}{2}H(X|Y = 1), \tag{24}$$

where $H(X|Y=i)$ stands for the entropy of X on the stratum $Y=i$. We notice that $H(X|Y=i) = H(Z)$, as the distribution of X on stratum $Y=i$ is normal with covariance matrix Σ, and its entropy does not depend on the mean. □

We note that, in Reference [17], entropy of one-dimensional Gaussian mixture $2^{-1}(N(a,1) + N(-a,1))$ is calculated as $h_e(a)$, where $h_e(a)$ is given in an integral form. As the entropy is invariant with respect to translation, function $h(a)$ defined above equals $h_e(a/2)$. The behavior of h and its two first derivatives is shown in Figure 1. It indicates that the function h is strictly increasing, and this fact is also stated in Reference [17] without proof. This is proved formally below. Strict monotonicity of h plays a crucial role in determining the order in which variables are included in a set of active variables. Note that $h(0) = \log(2\pi e)/2$, which is the entropy of the standard normal $N(0,1)$ variable. Values of h need to be calculated numerically.

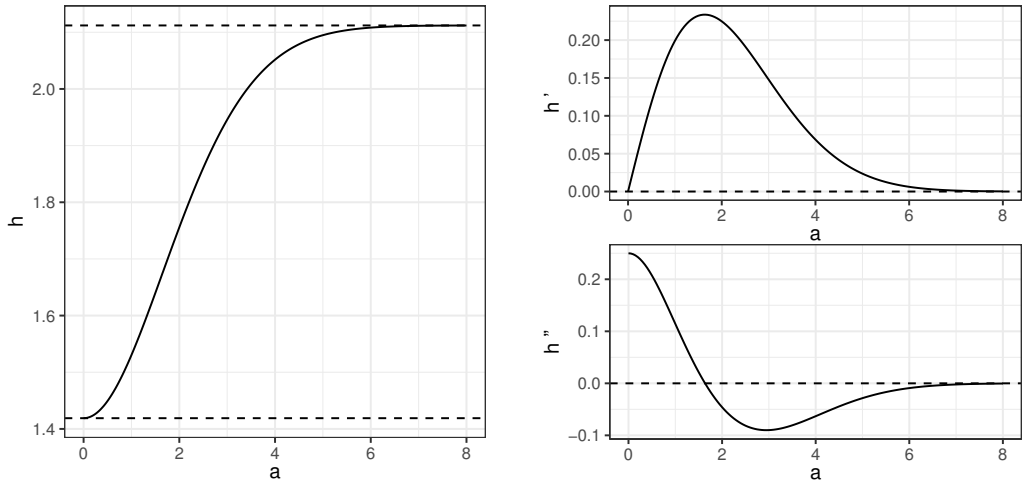

Figure 1. Behavior of function h and its two first derivatives. Horizontal lines in the left chart correspond to bounds of h and equal $\frac{1}{2}\log(2\pi e)$ and $\frac{1}{2}\log(2\pi e) + \log(2)$, respectively.

Lemma 1. *Differential entropy $h(a)$ of gaussian mixture defined in Theorem 1 is strictly increasing function of a.*

Proof. It is easy to see that h is differentiable and for calculation of its derivative, integration in (21) and taking derivatives can be interchanged. We show that derivative of h is positive. We have by standard manipulations, using the fact that $x\exp(-x^2/2)$ is an odd function for the second equality below, that

$$
\begin{aligned}
h'(a) &= -\frac{1}{2\sqrt{2\pi}}\int\limits_{\mathbb{R}}\left((x-a)e^{-\frac{(x-a)^2}{2}}\log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x-a)^2}{2}}\right)\right)+(x-a)e^{-\frac{(x-a)^2}{2}}\right)dx\\
&= -\frac{1}{2\sqrt{2\pi}}\int\limits_{\mathbb{R}}(x-a)e^{-\frac{(x-a)^2}{2}}\log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x-a)^2}{2}}\right)\right)dx\\
&= -\frac{1}{2\sqrt{2\pi}}\int\limits_{\mathbb{R}}xe^{-\frac{x^2}{2}}\log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x+a)^2}{2}}\right)\right)dx\\
&= -\frac{1}{2\sqrt{2\pi}}\int\limits_{0}^{\infty}xe^{-\frac{x^2}{2}}\log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x+a)^2}{2}}\right)\right)dx\\
&\quad -\frac{1}{2\sqrt{2\pi}}\int\limits_{-\infty}^{0}xe^{-\frac{x^2}{2}}\log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x+a)^2}{2}}\right)\right)dx\\
&= \frac{1}{2\sqrt{2\pi}}\int\limits_{0}^{\infty}xe^{-\frac{x^2}{2}}\left(\log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x-a)^2}{2}}\right)\right)-\log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x+a)^2}{2}}\right)\right)\right)dx.
\end{aligned}
$$

We have used change of variables for the third and the fifth equality above. It follows from the last expression that $h'(a) > 0$ as $(x-a)^2 < (x+a)^2$ for $x > 0$ and $a > 0$, and, therefore, h is increasing. □

Remark 1. *Note that Theorems 2 and 3 in conjunction with Lemma 1 show that entropy of mixture of two gaussians with the same covariance matrix and its mutual information with mixing distribution is strictly increasing function of the norm* $\left\|\Sigma^{-1}\mu\right\|$. *In particular, for* $\Sigma = I$, *entropy increases as the distance between centers of two gaussians increases. In addition, it follows from (22) and* $I(X,Y) \geq 0$ *that* $h(s) \geq \log(2\pi e)/2$ *for any* $s \in \mathbb{R}$.

Remark 2. *We call a random variable* $X \in \mathbb{R}^d$ *a generalized mixture when there exist diffeomorphisms* $f_i : \mathbb{R} \to \mathbb{R}$ *such that* $(f_1(X_1), \ldots f_p(X_d)) \sim 2^{-1}(\mathcal{N}(0, I_d) + \mathcal{N}(\mu, I_d))$. *Then, it follows from Theorem 2 that, analogously to Reference [20], that total correlation of X (cf. Reference [21]) defined as* $T(X) = \sum_{i=1}^{d} H(X_i) - H(X)$ *equals for generalized mixture X*

$$TC(X) = \sum_{i=1}^{d} h(|\mu_i|) - h(||\mu||) + (1-d)\log(2\pi e)/2,$$

where $\mu = (\mu_1, \ldots, \mu_d)^T$.

4. Main Results: Behavior of Information-Based Criteria in Generative Tree Model

In the following, we define a special gaussian Generative Tree Model and investigate how greedy procedure based on (14), as well as its analogues when CMI is replaced by JMI and CIFE, behaves in this model. Theorem 22 proved in the previous section will yield explicit formulae for CMIs in this model, whereas strict monotonicity of function $h(\cdot)$ proved in Lemma 1 will be essential to compare values of $I(X_j, Y|X_S)$ for different candidates X_j.

4.1. Generative Tree Model

We will consider the Generative Tree Model with tree structure illustrated in the Figure 2. Data Generating Process described by this model yields the distribution of the random vector $(Y, X_1, \ldots, X_{k+1}, X_1^{(1)})$ such that:

$$Y \sim \text{Bern}\,(1/2)\,, \quad X_i|Y \sim \mathcal{N}\left(\gamma^{i-1}Y, 1\right) \text{ and } i \in \{1, 2, \ldots, k+1\}, \quad |X_1 \sim \mathcal{N}\,(X_1, 1)\,, \tag{25}$$

where $0 < \gamma \leq 1$ is the parameter. Thus, first the value $Y = 0, 1$ is generated with both values 0 and 1 having the same probability $1/2$; then, $X_1, \ldots X_{k+1}$ are generated as normal variables with the variance 1 and the mean equal to Y. Finally, once the value of X_1 is obtained, $X_1^{(1)}$ is generated from normal distribution with the variance 1 and the mean equal to X_1. Thus, in the sense specified above, $X_1, \ldots X_{k+1}$ are the children of Y and $X_1^{(1)}$ is the child of X_1. Parameter γ controls how difficult the problem of feature selection is. Namely, the smaller the parameter γ is, the less information X_i holds about Y for $i \in \{1, 2, \ldots, k+1\}$. We will refer to the model defined above as $\mathcal{M}_{k,\gamma}$. We denote by, abusing slightly the notation, $p(y, x_i), p(x_1, x_1^{(1)})$ bivariate densities and by $p(y), p(x_i), p(x_1^{(1)})$ marginal densities. With this notation, the joint density $p(y, x_1, \ldots, x_{k+1}, x_1^{(1)})$ equals

$$p(y)\Big[\prod_{i=1}^{k+1}\frac{p(y,x_i)}{p(y)}\Big]\frac{p(x_1,x_1^{(1)})}{p(x_1)} = \frac{p(x_1,x_1^{(1)})}{p(x_1)p(x_1^{(1)})}\prod_{i=1}^{k+1}\frac{p(y,x_i)}{p(y)p(x_i)}\Big[\prod_{i=1}^{k+1}p(x_i)\Big]p(y)p(x_1^{(1)}),$$

which can be more succinctly written as

$$\prod_{(i,j)\in E}\frac{p(z_i,z_j)}{p(z_i)p(z_j)}\prod_{i\in V}p(z_i),$$

after renaming the variables to z_i, $i = 1, \ldots k+3$ and E and V standing for edges and vertices in the graph shown in Figure 2 (cf. formula (4.1) in Reference [4]).

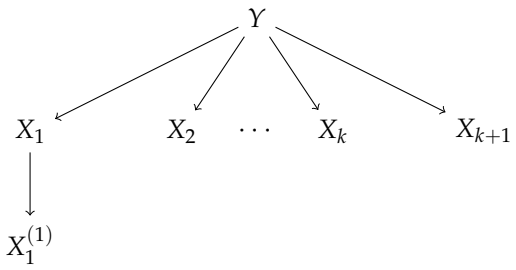

Figure 2. Generative Tree Model under consideration.

The above model generalizes the model discussed in Reference [3], but some branches which are irrelevant in our considerations are omitted. The values of conditional mutual information $I(X_{k+1}, Y|X_S)$ in the model, where $S = \{1, 2, \ldots, k\}$ for different γ as a function of k are shown in the Figure 3. We prove in the following that $I(X_{k+1}, Y|X_S) > 0$; thus, X_{k+1} carries non-null predictive information about Y even when variables $X_1, \ldots, X_k$ are already chosen as predictors. We note that $I(X_1^{(1)}, Y|X_S) = 0$ for every $\gamma \in (0, 1]$ and X_S containing X_1. Thus, $\{X_1, \ldots, X_{k+1}\}$ is the Markov Blanket (cf., e.g., Reference [22]) of Y among predictors $\{X_1, \ldots, X_{k+1}, X_1^{(1)}\}$ and $\{X_1, \ldots, X_{k+1}\}$ is sufficient for Y (cf. Reference [23]). A more general model may be considered which incorporates children of every vertex $X_1, \ldots, X_{k+1}$, and several levels of progeny. Here, we show how one variable $X_1^{(1)}$ which does not belong to Markov Blanket of Y is treated differently by the considered selection rules.

Intuitively, for $0 < \gamma < 1$ and $l < n$ X_l carry more information about Y than X_n and, moreover, $X_1^{(1)}$ is redundant once X_1 has been chosen. Thus, predictors should be chosen in order $X_1, X_2, \ldots X_{k+1}$. For $\gamma = 1$, the order of selection of X_i is also $X_1, \ldots, X_{k+1}$ in concordance with our convention of breaking ties, but $X_1^{(1)}$ should not be chosen. We show in the following that CMI chooses variables in this order; however, the order with respect to its approximations, CIFE, and JMI may be different. We also note that alternative way of representing predictors is

$$X_i = \gamma^{i-1} Y + \varepsilon_i, \qquad X_1^{(1)} = X_1 + \varepsilon_{k+2}, \tag{26}$$

for $i = 1, \ldots, k+1$, where $\varepsilon_1, \ldots, \varepsilon_{k+2}$ are i.i.d. $N(0,1)$. Thus, in particular

$$a_k Y = \sum_{i=1}^{k+1} X_i - \sum_{i=1}^{k+1} \varepsilon_i,$$

with $a_k = (1-\gamma^{k+1})/(1-\gamma)$. Moreover, it is seen that $\mathbb{E}X_i = \gamma^{i-1}\mathbb{E}Y = \gamma^{i-1}/2$.

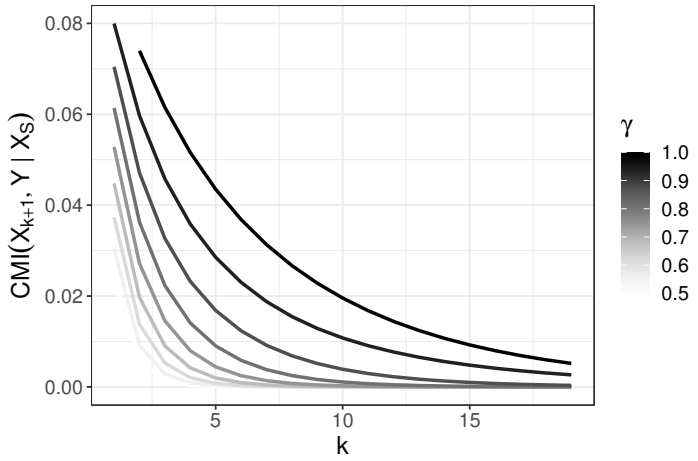

Figure 3. Behavior of conditional mutual information $I(X_{k+1}, Y|X_1, X_2, \ldots, X_k)$ as a function of k for different γ values.

It is shown in Reference [2] that maximization of $I(X_j, Y|X_S)$ is equivalent to maximization of $CIFE(X_j, Y|X_S)$ provided that selected features in X_S are independent and class-conditionally independent given unselected features X_j. It is easily seen that these properties do not hold in the considered GTM for $S = \{1, \ldots, l\}$ and $j = l+1$ for $l \leq k$. It can also be seen by a direct calculation that CMI differs from CIFE in GTM. Take $S = \{1, 2\}$ and $X_j = X_1^{(1)}$. Then, note that the difference between these quantities equals

$$I(X_j, Y|X_S) - I(X_j, Y) - \sum_{i \in S} II(X_i, X_j, Y) \tag{27}$$

Moreover, using conditional independence, we have

$$II(X_1, X_1^{(1)}, Y) = I(X_1^{(1)}, Y|X_1) - I(X_1^{(1)}, Y) = -I(X_1^{(1)}, Y)$$

and

$$II(X_2, X_1^{(1)}, Y) = I(X_1^{(1)}, X_2|Y) - I(X_1^{(1)}, X_2) = -I(X_1^{(1)}, X_2);$$

thus, plugging the above equalities into (27)and using $I(X_1^{(1)}, Y|X_1, X_2) = 0$, we obtain that expression there equals $I(X_1^{(1)}, X_2)$, which is strictly positive in the considered GTM.

Similar considerations concerning conditions stated above (18) show that maximization of JMI is not equivalent to maximization of CMI in GTM. Namely, if $S = \{1, 2\}$ and $j \in \{3, \ldots, k+1\}$, then it is easily seen that $I(X_j, X_{S\setminus\{i\}}|X_i) > 0$ and $I(X_j, X_{S\setminus\{i\}}|X_i, Y) = 0$ for $i = 1, 2$; thus, the last term in (17) is negative.

In order to support this numerically for a specific case, consider $\gamma = 2/3$. In the first column of the Table 1a, MI values $I(X_i, Y), i = 1, \ldots, 4$ are shown for this value of γ. They were calculated in Reference [3] using simulations, while here they are based on (23) and numerical evaluation of $h\left(\left\|\Sigma^{-1/2}\mu\right\|\right)$. Additionally, in Table 1, CMI values from subsequent steps and JMI and CIFE values in such a model are shown. As a foretaste of the analysis which follows, note that, in view of panel (b) of the table, JMI chooses erroneously $X_1^{(1)}$ in the third step instead of X_3 in contrast to CIFE (cf. part (c) of the table) which chooses X_1, X_2, X_3 in the right order. Note also that, in this case, is the second largest mutual informations with Y; thus, when the filter based solely on this information is considered, then $X_1^{(1)}$ is chosen at the second step (after X_1).

We note that analysis of behavior of CMI and its approximations including CIFE and JMI has been given in Reference [24], Section 6, for a simple model containing 4 predictors. We analyze here the behavior of these measures of conditional dependence for the general model $\mathcal{M}_{k,\gamma}$, which involves arbitrary number of predictors having varying dependence with Y.

Table 1. The criteria (Conditional Mutual Information (CMI), Joint Mutual Information (JMI), Conditional Infomax Feature Extraction (CIFE)) values for $k = 2$ and $\gamma = 2/3$. A value of the chosen variable in each step and for each criterion is in bold.

(a) $X_{S_1} = \{X_1\}, X_{S_2} = \{X_1, X_2\}, X_{S_3} = \{X_1, X_2, X_3\}$

	$I(\cdot, Y)$	$I(\cdot, Y\|X_{S_1})$	$I(\cdot, Y\|X_{S_2})$	$I(\cdot, Y\|X_{S_3})$
X_1	**0.1114**			
X_2	0.0527	**0.0422**		
X_3	0.0241	0.0192	**0.0176**	
$X_1^{(1)}$	0.0589	0.0000	0.0000	**0.0000**

(b) $X_{S_1} = \{X_1\}, X_{S_2} = \{X_1, X_2\}, X_{S_3} = \{X_1, X_2, X_1^{(1)}\}$

	$JMI(\cdot)$	$JMI(\cdot\|X_{S_1})$	$JMI(\cdot\|X_{S_2})$	$JMI(\cdot\|X_{S_3})$
X_1	**0.1114**			
X_2	0.0527	**0.0422**		
X_3	0.0241	0.0192	0.0205	**0.0208**
$X_1^{(1)}$	0.0589	0.0000	**0.0266**	

(c) $X_{S_1} = \{X_1\}, X_{S_2} = \{X_1, X_2\}, X_{S_3} = \{X_1, X_2, X_3\}$

	$CIFE(\cdot)$	$CIFE(\cdot\|X_{S_1})$	$CIFE(\cdot\|X_{S_2})$	$CIFE(\cdot\|X_{S_3})$
X_1	**0.1114**			
X_2	0.0527	**0.0422**		
X_3	0.0241	0.0192	**0.0169**	
$X_1^{(1)}$	0.0589	0.0000	−0.0057	**−0.0083**

4.2. Behavior of CMI

First of all, we show that the criterion based on conditional mutual information CMI without any modifications chooses correct variables in the right order. It has been previously noticed that $I(X_1^{(1)}, Y|X_S) = 0$ for $S = \{1, \ldots, k\}$. Now, we show that $I(X_{k+1}, Y|X_S) > 0$ for every k. Namely, applying Theorem 3 and the chain rule for mutual information

$$I(X_{S\cup\{k+1\}}, Y) = I(X_S, Y) + I(X_{k+1}, Y|X_S),$$

we obtain

$$I(X_{k+1}, Y|X_S) = h\left(\sqrt{\sum_{i=0}^{k}\gamma^{2i}}\right) - h\left(\sqrt{\sum_{i=0}^{k-1}\gamma^{2i}}\right) > 0, \tag{28}$$

where the inequality follows as h is an strictly increasing function. Thus, we proved that $I(X_1^{(1)}, Y|X_S) = 0 < I(X_{k+1}, Y|X_S)$ for $S = \{1, \ldots, k\}$ for every k. Whence we have for $S = \{1, \ldots, l\}$ and $l < k$ that

$$\arg\max_{Z \in S^c} I(Z, Y|X_S) = X_{l+1},$$

thus CMI chooses predictors in a correct order. Figure 3 shows behavior of $g(k, \gamma) = I(X_{k+1}, Y|X_1, \ldots, X_k)$ as the function of k for various γ. Note that it follows from Figure 3 that $g(\cdot, \gamma)$ is decreasing. This means that the additional information on Y obtained when X_{k+1} is incorporated gets smaller with k. Now, we study the order in which predictors are chosen with respect to JMI and CIFE.

4.3. Behavior of JMI

The main objective of this section is to examine performance of JMI criterion in the Generative Tree Model for different values of parameter γ. We will show that:

- For $\gamma = 1$ active predictors $X_1, \ldots, X_{k+1} \in MB(Y)$ are chosen in the right order and $X_1^{(1)}$ is not chosen before them;
- For $0 < \gamma < 1$, variable $X_1^{(1)} \notin MB(Y)$ is chosen at a certain step before all $X_1, \ldots, X_{k+1}$ are chosen, and we evaluate a moment when this situation occurs.

Consider the model above and assume that the set of indices of currently chosen variables equals $S = \{1, 2, \ldots, k\}$. For $i \in \{1, 2, \ldots, k\}$ we apply chain rule (6) and Theorem 3 with the following covariance matrices and mean vectors for $I((X_i, Z), Y)$ (cf. (26)):

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mu = \begin{pmatrix} \gamma^{i-1} \\ \gamma^k \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \mu = \begin{pmatrix} \gamma^{i-1} \\ 1 \end{pmatrix}, \tag{29}$$

respectively, for $Z = X_{k+1}$ and $Z = X_1^{(1)}$. Then, we have

$$I(X_{k+1}, Y|X_i) = h\left(\sqrt{\gamma^{2k} + \gamma^{2(i-1)}}\right) - h\left(\gamma^{i-1}\right), \tag{30}$$

$$I(X_1^{(1)}, Y|X_i) = h\left(\sqrt{\gamma^{2(i-1)} + \frac{1}{2}}\right) - h\left(\gamma^{i-1}\right) \text{ for } i \neq 1, \tag{31}$$

$$I(X_1^{(1)}, Y|X_1) = 0. \tag{32}$$

The last equation follows from the fact that $X_1^{(1)}$ and Y are conditionally independent given X_1.

From the definition of $JMI(X, Y|X_S)$, abbreviated from now on to $JMI(X|X_S)$ to simplify notation, we obtain

$$kJMI(X_{k+1}|X_S) = \sum_{i=1}^{k}\left(h\left(\sqrt{\gamma^{2k} + \gamma^{2(i-1)}}\right) - h\left(\gamma^{i-1}\right)\right), \tag{33}$$

$$kJMI(X_1^{(1)}|X_S) = \begin{cases} 0 & \text{if } k = 1 \\ \sum\limits_{i=2}^{k}\left(h\left(\sqrt{\gamma^{2(i-1)} + \frac{1}{2}}\right) - h\left(\gamma^{i-1}\right)\right) & \text{if } k > 1 \end{cases}. \tag{34}$$

We observe that the variables $X_1, X_2, \ldots$ are chosen in order according to JMI, as for $S = \{1, \ldots, l\}$ and $l < m < n$, we have $JMI(X_m) > JMI(X_n)$. For $\gamma = 1$, the right-hand sides of the last two

expressions equal $k\left(h\left(\sqrt{2}\right)-h\left(1\right)\right)$ and $(k-1)\left(h\left(\sqrt{3/2}\right)-h\left(1\right)\right)$, respectively. Thus, for $\gamma=1$, we have $JMI(X_{k+1}|X_S) > JMI(X_1^{(1)}|X_S)$, which means that variables are chosen in the order $X_1,\ldots,X_{k+1}$ and $X_1^{(1)}$ is not chosen before them when JMI criterion is used. Although, for $\gamma=1$, JMI criterion does not select this redundant feature, we note that, for $k\to\infty$, $S=\{1,\ldots,k\}$, and $\gamma=1$

$$JMI(X_1^{(1)}|X_S)\to\left(h\left(\sqrt{\frac{3}{2}}\right)-h\left(1\right)\right)>0,$$

which differs from $I(X_1^{(1)},Y|X_S)=0$ for all $k\geq 1$. We note also that, in this case, $JMI(X_{k+1}|X_S)$ does not depend on k in contrast to $I(X_{k+1},Y|X_S)$.

Now, we will consider the case $0<\gamma<1$. We want to show that, for sufficiently large k and $S=\{1,\ldots,k\}$, JMI criterion chooses $X_1^{(1)}$ since

$$JMI(X_{k+1}|X_S) < JMI(X_1^{(1)}|X_S).$$

The last inequality is equivalent to

$$\sum_{i=2}^{k}\left(h\left(\sqrt{\gamma^{2(i-1)}+\frac{1}{2}}\right)-h\left(\sqrt{\gamma^{2k}+\gamma^{2(i-1)}}\right)\right)>h(\sqrt{1+\gamma^{2k}})-h\left(1\right). \tag{35}$$

The right-hand side tends to 0 when $k\to\infty$. For the left-hand side, note that, for $k>-\frac{\log_\gamma 2}{2}$, we have $\gamma^{2k}<1/2$, and all summands of the sum above are positive, as h is an increasing function. Thus, bounding the sum by its first term, we have

$$\sum_{i=2}^{k}\left(h\left(\sqrt{\gamma^{2(i-1)}+\frac{1}{2}}\right)-h\left(\sqrt{\gamma^{2k}+\gamma^{2(i-1)}}\right)\right)>h(\sqrt{\gamma^2+1/2})-h(\sqrt{\gamma^2+1/2})=0.$$

The minimal k for which the JMI criterion incorrectly chooses $X_1^{(1)}$, i.e., the first k for which (35) holds, is shown in Figure 4. The values of JMI criterion for variables X_{k+1} and $X_1^{(1)}$ is shown in Figure 5. Figure 4 indicates that $X_1^{(1)}$ is chosen early; for $\gamma\leq 0.8$, it happens in the third step at the latest.

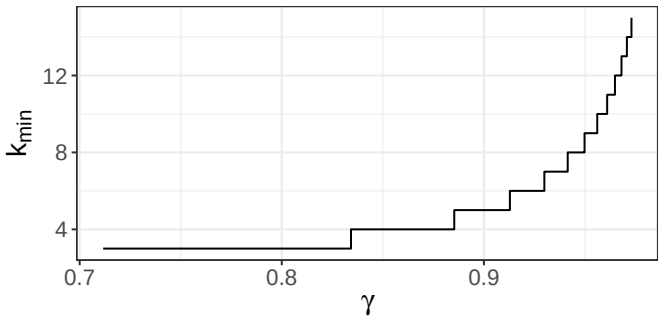

Figure 4. Minimal k for which $JMI(X_{k+1}|X_S) < JMI(X_1^{(1)}|X_S)$, $0<\gamma<1$.

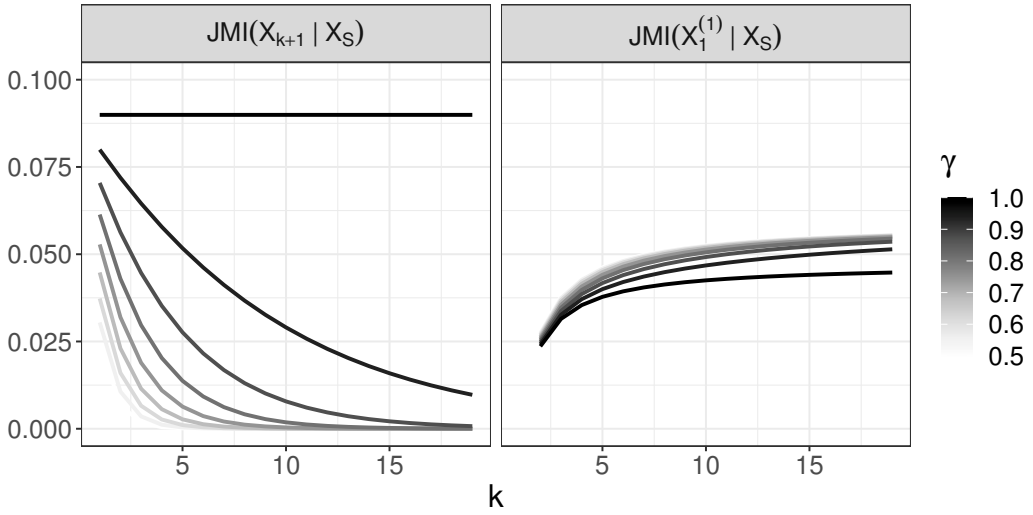

Figure 5. The behavior of JMI in the generative tree model: $JMI(X_{k+1}|X_S)$ and $JMI(X_1^{(1)}|X_S)$.

4.4. Behavior of CIFE and Its Comparison With JMI

The aim of this section is to show that, although both JMI and CIFE criteria are developed as approximations to conditional mutual information, their behavior in the tree generative model differs. We will show that:

- For $\gamma = 1$, CIFE incorrectly chooses $X_1^{(1)}$ at some point;
- For $0 < \gamma < 1$, CIFE selects variables $X_1, \ldots, X_{k+1}$ in the right order.

Thus, CIFE behaves very differently from JMI in Generative Tree Model.
Analogously to formulae for JMI, we have the following formulae for CIFE ($S = \{1, \ldots, k\}$):

$$CIFE(X_{k+1}|X_S) = (1-k)\left(h\left(\gamma^k\right) - \frac{1}{2}\log(2\pi e)\right) + \sum_{i=1}^{k}\left(h\left(\sqrt{\gamma^{2k} + \gamma^{2(i-1)}}\right) - h\left(\gamma^{i-1}\right)\right),$$

$$CIFE(X_1^{(1)}|X_S) = \begin{cases} 0 & \text{if } k = 1 \\ (1-k)\left(h(1) - \frac{1}{2}\log(2\pi e)\right) + \sum_{i=2}^{k}\left(h\left(\sqrt{\gamma^{2(i-1)} + \frac{1}{2}}\right) - h\left(\gamma^{i-1}\right)\right) & \text{if } k > 1 \end{cases}.$$

For $\gamma = 1$, we have

$$\begin{aligned} CIFE(X_{k+1}|X_S) &= (1-k)\left(h(1) - \frac{1}{2}\log(2\pi e)\right) + \sum_{i=1}^{k}\left(h\left(\sqrt{2}\right) - h(1)\right), \\ &= h(1) - \frac{1}{2}\log(2\pi e) - k\left(2h(1) - h(\sqrt{2}) - \frac{1}{2}\log(2\pi e)\right) \\ CIFE(X_1^{(1)}|X_S) &= (1-k)\left(2h(1) - \frac{1}{2}\log(2\pi e) - h\left(\sqrt{\frac{3}{2}}\right)\right). \end{aligned}$$

Note that both expressions above are linear functions with respect to k. Comparison of their slopes, in view of $h\left(\sqrt{\frac{3}{2}}\right) < h\left(\sqrt{2}\right)$ as h is an increasing function, yields that, for sufficiently large k, we obtain $CIFE(X_{k+1}|X_S) < CIFE(X_1^{(1)}|X_S)$. The behavior of CIFE for $0 < \gamma < 1$ in case of X_{k+1} and $X_1^{(1)}$ is shown in Figure 6 and the difference between $CIFE(X_{k+1}|X_S)$ and $CIFE(X_1^{(1)}|X_S)$ in Figure 7. The values below 0 in the last plot occur for $\gamma = 1$; only, thus, for $0 < \gamma < 1$, we have $CIFE(X_{k+1}|X_S) > CIFE(X_1^{(1)}|X_S)$ for any k.

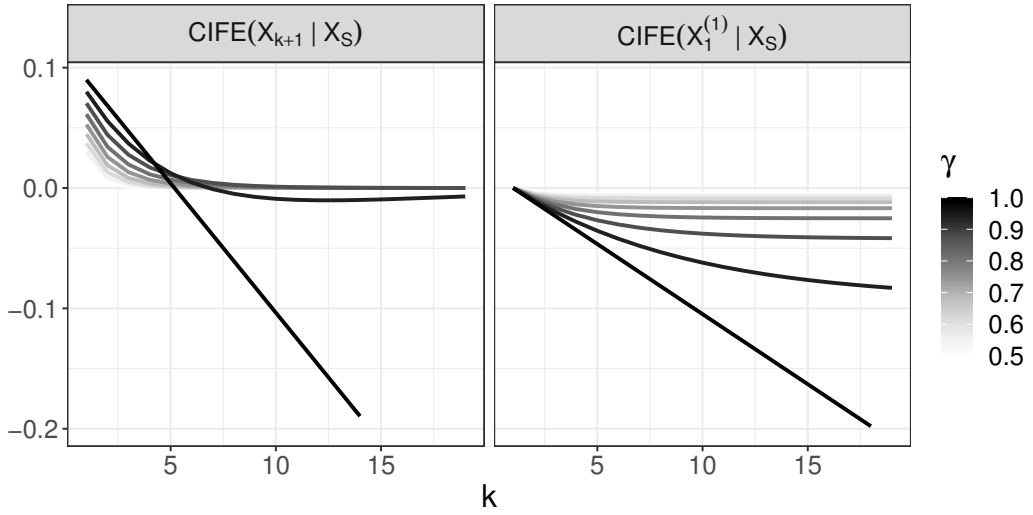

Figure 6. The behavior of CIFE in the generative tree model: $CIFE(X_{k+1}|X_S)$ and $CIFE(X_1^{(1)}|X_S)$.

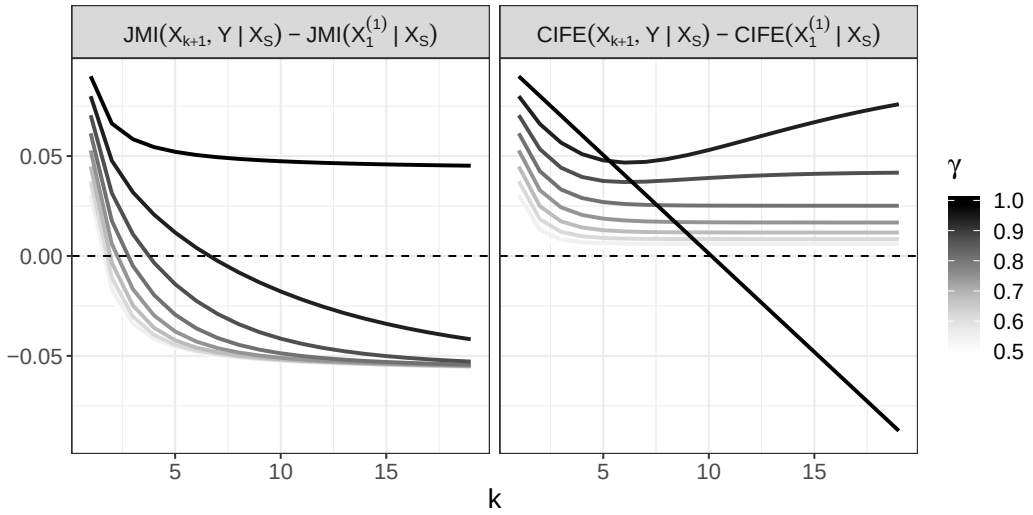

Figure 7. Difference between values of JMI for X_{k+1} and $X_1^{(1)}$ (**left panel**) and analogous difference for CIFE (**right panel**). Values below 0 mean that the variable $X_1^{(1)}$ is chosen.

Furthermore, as $2h(1) - \frac{1}{2}\log(2\pi e) - h\left(\sqrt{\frac{3}{2}}\right) \approx 0.0642 > 0$, we have, for $\gamma = 1$,

$$CIFE(X_1^{(1)}|X_S) \to -\infty \text{ as } k \to \infty,$$

and as $2h(1) - h(\sqrt{2}) - \frac{1}{2}\log(2\pi e) \approx 0.0215 > 0$, we have

$$CIFE(X_{k+1}|X_S) \to -\infty \text{ as } k \to \infty.$$

In order to understand the consequences of this property, let us momentarily assume that one introduces an intuitive stopping rule which says that candidate X_{j_0} such that $j_0 = \arg\max_{j \in S^c} CIFE(X_j, Y|X_S)$ is appended only when $CIFE(X_{j_0}, Y|X_S) > 0$. Then, Positive Selection Rate (PSR) of such selection procedure may become arbitrarily small in model $\mathcal{M}_{k,\gamma}$ for fixed γ and sufficiently large k. PSR is defined as $|\hat{t} \cap t|/|t|$, where $t = \{1, \ldots, k+1\}$ is a set of indices of Markov Blanket of Y and $\hat{t}$ is a set of indices of the chosen variables.

5. Conclusions

We have considered $\mathcal{M}_{k,\gamma}$, a special case of Generative Tree Model and investigated behavior of CMI and related criteria JMI and CIFE in this model. We have shown that, despite the fact that both of these criteria are derived as approximations of CMI under certain dependence conditions, their behavior may greatly differ from that of CMI in the sense that they may switch the order of variable importance and treat inactive variables as more relevant than active ones. In particular, this occurs for JMI when $\gamma < 1$ and CIFE for $\gamma = 1$. We have also shown a drawback of CIFE procedure which consists in disregarding significant part of active variables so that PSR may become arbitrarily small in model $\mathcal{M}_{k,\gamma}$ for large k. As a byproduct, we obtain formulae for the entropy of multivariate gaussian mixture and its mutual information with mixing variable. We have also shown that the entropy of the gaussian mixture is a strictly increasing function of the euclidean distance between two centers of its components. Note that, in this paper, we investigated behavior of theoretical CMI and its approximations in GTM; for their empirical versions, we may expect exacerbation of effects described here.

Author Contributions: Conceptualization, M.Ł.; Formal analysis, J.M. and M.Ł.; Methodology, J.M. and M.Ł.; Supervision, J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: Comments of two referees which helped to improve presentation of the original version of the manuscript are gratefully acknowledged.

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

References

1. Guyon, I.; Elyseeff, A. An introduction to feature selection. In *Feature Extraction, Foundations and Applications*, Springer: Berlin/Heidelberger, Germany, 2006; Volume 207, pp. 1–25.
2. Brown, G.; Pocock, A.; Zhao, M.; Luján, M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* **2012**, *13*, 27–66.
3. Gao, S.; Ver Steeg, G.; Galstyan, A. Variational Information Maximization for Feature Selection. In *Advances in neural information processing systems*, MIT Press: Cambridge, MA, USA, 2016; pp. 487–495.
4. Lafferty, J.; Liu, H.; Wasserman, L. parse nonparametric graphical models. *Stat. Sci.* **2012**, *27*, 519–537.
5. Liu, H.; Xu, M.; Gu, H.; Gupta, A.; Lafferty, J.; Wasserman, L. Forest density estimation. *J. Mach. Learn. Res.* **2011**, *12*, 907–951.
6. Cover, T.M.; Thomas, J.A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*; Wiley-VCH: Hoboken, NJ, USA, 2006.
7. Yeung, R.W. *A First Course in Information Theory*; Kluwer: South Holland, Netherlands, 2002.
8. McGill, W.J. Multivariate information transmission. *Psychometrika* **1954**, *19*, 97–116.
9. Ting, H.K. On the Amount of Information. *Theory Probab. Appl.* **1960**, *7*, 439–447.
10. Han, T.S. Multiple mutual informations and multiple interactions in frequency data. *Inform. Control* **1980**, *46*, 26 – 45.
11. 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.
12. Vergara, J.R.; Estévez, P.A. A review of feature selection methods based on mutual information. *Neural. Comput. Appl.* **2014**, *24*, 175–186.
13. Lin, D.; Tang, X. Conditional infomax learning: an integrated framework for feature extraction and fusion. In *European conference on computer vision 2006 May 7*, Springer: Berlin/Heidelberger, Germany, 2006; pp. 68–82.
14. Mielniczuk, J.; Teisseyre, P. Stopping rules for information-based feature selection. *Neurocomputing* **2019**, *358*, 255–274.
15. Yang, H.H.; Moody, J. Data visualization and feature selection: new algorithms for nongaussian data. *Adv. Neural. Inf. Process Syst.* **1999**, *12*, 687–693.

16. Peng, H.; L., F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238.
17. Michalowicz, J.; Nichols, J.M.; Bucholtz, F. Calculation of differential entropy for a mixed gaussian distribution. *Entropy* **2008**, *10*, 200–206.
18. Moshkar, K.; Khandani, A. Arbitrarily tight bound on differential entropy of gaussian mixtures. *IEEE Trans. Inf. Theory* **2016**, *62*, 3340–3354.
19. Huber, M.; Bailey, T.; Durrant-Whyte, H.; Hanebeck, U. On entropy approximation for gaussian mixture random vectors. In *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, IEEE: Piscataway, NJ, USA, 2008; pp. 181–189.
20. Singh, S.; Póczos, B. Nonparanormal information estimation. *arXiv* **2017**, arXiv: 1702.07803.
21. Watanabe, S. Iformation theoretical analysis of multivariate correlation. *IBM J. Res. Dev.* **1960**, *45*, 211 – 232.
22. Pena, J.M.; Nilsson, R.; Bjoerkegren, J.; Tegner, J. Towards scalable and data efficient learning of Markov boundaries. *Int. J. Approx. Reasoning* **2007**, *45*, 211 – 232.
23. Achille, A.; Soatto, S. Emergence of invariance and disentanglements in deep representations. *J. Mach. Learn. Res.* **2018**, *19*, 1948–1980.
24. Macedo, F.; Oliveira, M.; Pacecho, A.; Valadas, R. Theoretical foundations of forward feature selection based on mutual information. *Neurocomputing* **2019**, *325*, 67–89.