Revisiting Probability Distribution Assumptions for Information Theoretic Feature Selection

Yuan Sun,[1*] **Wei Wang,**[2] **Michael Kirley,**[2] **Xiaodong Li,**[1] **Jeffrey Chan**[1]
[1]RMIT University, Melbourne, Australia
[2]University of Melbourne, Parkville, Australia
[1]{yuan.sun, xiaodong.li, jeffrey.chan}@rmit.edu.au
[2]weiw8@student.unimelb.edu.au, mkirley@unimelb.edu.au

Abstract

Feature selection has been shown to be beneficial for many data mining and machine learning tasks, especially for big data analytics. Mutual Information (MI) is a well-known information-theoretic approach used to evaluate the relevance of feature subsets and class labels. However, estimating high-dimensional MI poses significant challenges. Consequently, a great deal of research has focused on using low-order MI approximations or computing a lower bound on MI called Variational Information (VI). These methods often require certain assumptions made on the probability distributions of features such that these distributions are realistic yet tractable to compute. In this paper, we reveal two sets of distribution assumptions underlying many MI and VI based methods: *Feature Independence Distribution* and *Geometric Mean Distribution*. We systematically analyse their strengths and weaknesses and propose a logical extension called *Arithmetic Mean Distribution*, which leads to an unbiased and normalised estimation of probability densities. We conduct detailed empirical studies across a suite of 29 real-world classification problems and illustrate improved prediction accuracy of our methods based on the identification of more informative features, thus providing support for our theoretical findings.

1 Introduction

In the era of big data, we are often confronted with machine learning tasks involving a large number of features (Li et al. 2017; Bolón-Canedo, Sánchez-Maroño, and Alonso-Betanzos 2015). In many real-world applications, it is possible to improve the predictability, interpretability and training efficiency of specific machine learning models, including deep learning models (Gao, Ver Steeg, and Galstyan 2016; Cai et al. 2018), by selecting a subset of features for processing and/or by removing the irrelevant or redundant features (Bagherzadeh-Khiabani et al. 2016; Pascoal et al. 2012; Saeys, Inza, and Larrañaga 2007).

Feature selection methods can be divided into three categories: *wrapper*, *embedded* and *filter* (Guyon and Elisseeff 2003; Xue et al. 2016; Vergara and Estévez 2014). *Wrapper* and *embedded* methods rely on a classifier, using classification accuracy as an indication of feature quality. In contrast, *filter* methods employ an objective function, which measures the relevance between features and class labels. Perhaps the most widely used information-theoretic measure for objective function design is based on Mutual Information (MI) (see (Vergara and Estévez 2014; Li et al. 2017) for a comprehensive review and (Brown et al. 2012) for a unifying framework). However, it is a significant challenge to estimate high-dimensional MI from a limited number of samples (Wolpert and Wolf 1995; Brown et al. 2012). Consequently, many information-theoretic methods use low-dimensional MI approximations which necessitates specific assumptions to be made on the probability distribution of features (Balagani and Phoha 2010; Brown et al. 2012; Vinh et al. 2016; Gao, Ver Steeg, and Galstyan 2016). An alternative way to design objective function for feature selection is based on a lower bound on MI called Variational Information (VI) (Gao, Ver Steeg, and Galstyan 2016). This lower bound is tractable to compute if a variational probability distribution of features is carefully chosen, and it becomes exact when the variational distribution matches the real distribution of features. Thus the assumption on variational probability distribution of features also has a key role in VI based methods.

In this paper, we re-examine the assumptions on the probability distributions of features used in typical information-theoretic feature selection methods. Our overarching research goal is to answer the following questions: *"Which distribution of features in a given information-theoretic feature selection method is both realistic and tractable to compute?"* and *"How do particular assumptions related to the probability distribution of features impact on classification accuracy?"* Importantly, we will show that many information-theoretic feature selection methods in literature differ only in their underlying assumptions on feature distributions. Specifically, we systematically analyse three sets of assumptions on feature distributions: *Feature Independence Distribution* (FID), *Geometric Mean Distribution* (GMD) and *Arithmetic Mean Distribution* (AMD).

- The FID assumption has been widely accepted in literature. However we show that the probability density estimation under the FID assumption has certain bias.

*Corresponding author

- The GMD assumption relaxes FID and has potential to reduce the estimation bias. However we show that the probability density estimates under GMD may not be normalized, i.e., the probability density integration is not one.
- The AMD assumption is a logical extension to fix the normalization issue inherent in the GMD assumption.

To demonstrate the efficacy of assumptions related to the probability distributions of features, we report classification results across a suite of 29 real-world datasets, where each of the distribution assumptions was incorporated into the MI and VI frameworks. The experimental results show that the AMD assumption typically leads to better classification accuracy than both the FID and GMD assumptions.

2 Mutual Information Based Methods

Given a supervised learning task with feature vector $\boldsymbol{X}$ and class label C, the goal of MI based feature selection methods is to search for a subset of K features ($\boldsymbol{S}$) such that the MI between $\boldsymbol{S}$ and C is maximized:

$$\boldsymbol{S} = \underset{\boldsymbol{S}\subset\boldsymbol{X},|\boldsymbol{S}|=K}{\arg\max}\ I(\boldsymbol{S};C), \tag{1}$$

where $I(\cdot)$ denotes MI (Cover and Thomas 2012).

This search problem can be modeled as quadratic programming problem and solved globally (Rodriguez-Lujan et al. 2010; Nguyen et al. 2014; Venkateswara et al. 2015). In contrast, Sequential Forward Selection (Kittler 1986; Pudil, Novovičova, and Kittler 1994) is a class of greedy methods that selects a candidate feature X_m at a time such that the MI between X_m and C given the selected feature subset ($\boldsymbol{S}$) is maximized:

$$X_m = \underset{X_m\in\boldsymbol{X}\setminus\boldsymbol{S}}{\arg\max}\ \big\{J(X_m) := I(X_m;C|\boldsymbol{S})\big\}, \tag{2}$$

where $\boldsymbol{X}\setminus\boldsymbol{S}$ denotes $\boldsymbol{X}$ excluding $\boldsymbol{S}$ (the unselected features), and $J(\cdot)$ denotes the objective function.

In practice, it is difficult to estimate the high-order conditional MI $I(X_m;C|\boldsymbol{S})$. A possible way to do this is using the *functional estimation*, where MI is directly estimated without computing probability densities (Wu and Yang 2016; Noshad, Zeng, and Hero 2019). However, in the feature selection community, a large number of methods has used low-order MI to approximate the high-order MI, including (Yang and Moody 1999; Peng, Long, and Ding 2005; Vinh et al. 2016; Wang et al. 2017; Zadeh et al. 2017; Singha and Shenoy 2018; Sharmin et al. 2019). These methods are typically based on the concept of feature relevancy, redundancy and complementarity (Brown et al. 2012; Vergara and Estévez 2014). To justify the low-order estimation, probability distribution assumptions such as feature independence and class-conditional feature independence have been used (Balagani and Phoha 2010; Brown et al. 2012; Vergara and Estévez 2014; Venkateswara et al. 2015; Vinh et al. 2016; Gao, Ver Steeg, and Galstyan 2016). In the following subsection, we will revisit these assumptions.

2.1 Methods Based on Feature Independence Distribution

Given selected features $\boldsymbol{S}$ and a candidate feature X_m, let $\boldsymbol{S}_k \subseteq \boldsymbol{S}$ contain k ($0 \le k \le |\boldsymbol{S}|$) randomly selected features from $\boldsymbol{S}$. Define a trial feature set $\boldsymbol{T} = X_m \cup \boldsymbol{S}$ and its subset $\boldsymbol{T}_k = X_m \cup \boldsymbol{S}_{k-1}$ where the first feature is X_m and the remaining $k-1$ features are from $\boldsymbol{S}$; $\boldsymbol{T}_0 = \emptyset$. The *feature independence distribution* assumption is presented as

Feature Independence Distribution (FID). The selected features $\boldsymbol{S}$ and a candidate feature X_m are independent and class-conditionally independent at order k ($0 \le k \le |\boldsymbol{S}|$):

$$p(\boldsymbol{T}\setminus\boldsymbol{T}_k|\boldsymbol{T}_k) \simeq \prod_{X_i\in\boldsymbol{T}\setminus\boldsymbol{T}_k} p(X_i|\boldsymbol{T}_k), \tag{3}$$

$$p(\boldsymbol{T}\setminus\boldsymbol{T}_k|\boldsymbol{T}_k, C) \simeq \prod_{X_i\in\boldsymbol{T}\setminus\boldsymbol{T}_k} p(X_i|\boldsymbol{T}_k, C). \tag{4}$$

We use $\simeq$ to denote "asymptotic" equality, in the sense that Eq. (3) and (4) will become exact when $k = |\boldsymbol{S}|$.

The hyper-parameter k controls the trade-off between the amount of information loss and model complexity. In theory, considering a larger k leads to a more realistic estimation for $p(\boldsymbol{T}\setminus\boldsymbol{T}_k|\boldsymbol{T}_k)$. But in practice, this is not guaranteed to be beneficial because estimating high-order probability density function is typically problematic and requires many samples and computational resources. Existing assumptions made on feature distribution are often up to order 2.

Note that FID is a generic case of several feature distribution assumptions used in literature. By simply setting $k = 0$, 1 or 2, we restore the assumptions used in (Balagani and Phoha 2010; Brown et al. 2012; Vinh et al. 2016). These assumptions are generally strong, and only hold when the given features are independent or conditionally independent.

Theorem 1. *Under the FID assumption of order k, the objective function in Eq. (2) is equivalent to:*

$$J^{\mathrm{k,k}}_{\mathrm{FID}}(X_m) \sim \sum_{i=1}^{k} I(X_{t_i};C|\boldsymbol{T}_{i-1}) + \sum_{X_i\in\boldsymbol{T}\setminus\boldsymbol{T}_k} I(X_i;C|\boldsymbol{T}_k), \tag{5}$$

where $\boldsymbol{T}_{i-1} = \{X_{t_1},\cdots,X_{t_{i-1}}\}$ and $X_{t_1} = X_m$; and $\sim$ denotes "equivalent to". More generally, we consider different values of k in Eq. (3) and (4), denoted as k_1 and k_2 respectively. The objective function $J(X_m)$ is equivalent to:

$$\begin{aligned} J^{\mathrm{k_1,k_2}}_{\mathrm{FID}}(X_m) \sim{}& \sum_{i=1}^{k_1} H(X_{t_i}|\boldsymbol{T}_{i-1}) + \sum_{X_i\in\boldsymbol{T}\setminus\boldsymbol{T}_{k_1}} H(X_i|\boldsymbol{T}_{k_1}) \\ &- \sum_{i=1}^{k_2} H(X_{t_i}|\boldsymbol{T}_{i-1},C) - \sum_{X_i\in\boldsymbol{T}\setminus\boldsymbol{T}_{k_2}} H(X_i|\boldsymbol{T}_{k_2},C), \end{aligned} \tag{6}$$

where $H(\cdot)$ denotes entropy (Cover and Thomas 2012).

In Theorem 1, we have obtained a set of methods based on the FID assumption with the objective function defined in Eq. (6). The proof of Theorem 1 is in Appendix I.A.

Existing Methods Using FID. By varying the values of k_1 and k_2, we can recover three existing MI based methods: Mutual Information Maximization (MIM) (Lewis 1992), Mutual Information Feature Selection (MIFS) (Battiti 1994) and Conditional Informative Feature Extraction (CIFE) (Lin and Tang 2006). Their objective functions are listed in Table 1 (J_{FID}), and the proof shown in Appendix I.A. *We have identified these three methods are all based on the FID assumption, and differ only in the order k used.*

Table 1: A brief summary of existing MI based methods that use the FID (J_{FID}) and GMD (J_{GMD}) assumptions.

J	k_1	k_2	Name	Objective Function	Reference
J_{FID}	0	0	MIM	$I(X_m;C)$	(Lewis 1992)
	1	0	MIFS	$I(X_m;C)-\beta\sum_{X_i\in\boldsymbol{S}} I(X_m;X_i)$ with $\beta=1$	(Battiti 1994)
	1	1	CIFE	$I(X_m;C)-\sum_{X_i\in\boldsymbol{S}} I(X_m;X_i)+\sum_{X_i\in\boldsymbol{S}} I(X_m;X_i\|C)$	(Lin and Tang 2006)
J_{GMD}	0	0	MIM	$I(X_m;C)$	(Lewis 1992)
	1	0	MRMR	$I(X_m;C)-1/\|\boldsymbol{S}\|\sum_{X_i\in\boldsymbol{S}} I(X_m;X_i)$	(Peng, Long, and Ding 2005)
	1	1	JMI	$I(X_m;C)-1/\|\boldsymbol{S}\|\sum_{X_i\in\boldsymbol{S}} I(X_m;X_i)+1/\|\boldsymbol{S}\|\sum_{X_i\in\boldsymbol{S}} I(X_m;X_i\|C)$	(Yang and Moody 1999)
	2	1	RMRMR	$I(X_m;C)-1/\|\boldsymbol{S}\|\sum_{X_i\in\boldsymbol{S}} I(X_m;X_i)+1/\|\boldsymbol{S}\|\sum_{X_i\in\boldsymbol{S}} I(X_m;X_i\|C)$ $-1/\|\boldsymbol{S}\|/(\|\boldsymbol{S}\|-1)\sum_{X_i\in\boldsymbol{S}}\sum_{X_j\in\boldsymbol{S}\setminus X_i} I(X_m;X_j\|X_i)$	(Vinh et al. 2016)

Time Complexity of FID Methods. We briefly analyze the time complexity of the methods presented in Theorem 1 with different k_1 and k_2 values ($J_{\mathrm{FID}}^{\mathrm{k_1,k_2}}$). Suppose N is the number of samples; M is the number of features. We assume the number of joint probability states in $p(X_i|\boldsymbol{T}_k;C)$ is less than the sample size (otherwise the sample size is insufficient), so that the joint entropy $H(X_i|\boldsymbol{T}_{\boldsymbol{k}},C)$ can be calculated in $\mathcal{O}(N)$ time. If the maximum number of features to be selected is K, the number of joint entropy values to be calculated is in $\mathcal{O}(KM)$. Thus, the time complexity of $J_{\mathrm{FID}}^{\mathrm{k_1,k_2}}$ is $\mathcal{O}(KMN)$ when $k_1\geq 1$ or $k_2\geq 1$. When $k_1=k_2=0$, the time complexity of $J_{\mathrm{FID}}^{0,0}$ is $\mathcal{O}(MN)$, as only M MI values need to be calculated.

Strength and Limitation of FID. The FID assumption is simple and easy to interpret. Under the FID assumption, the high-dimensional MI is tractable to compute as it can be directly decomposed into a series of low-order MI approximations. However, the FID assumption is strong and causes bias in the estimation of probability density, as shown subsequently. This assumption needs to specify a feature subset $\boldsymbol{T}_k$ as conditional features. As the order of features to be selected is not fixed, it is reasonable to assume that $\boldsymbol{T}_k$ is any subset of $\boldsymbol{T}$ (Venkateswara et al. 2015). Specifically we assume $\boldsymbol{T}_k=\boldsymbol{S}_k$, i.e., $\boldsymbol{T}_k$ is any k features from $\boldsymbol{S}$. A simple calculation yields $p(X_m|\boldsymbol{S})=\frac{p(\boldsymbol{T}\setminus\boldsymbol{S}_k|\boldsymbol{S}_k)}{p(\boldsymbol{S}\setminus\boldsymbol{S}_k|\boldsymbol{S}_k)}$. Using the assumption that features in $\boldsymbol{T}\setminus\boldsymbol{S}_k$ are independent given $\boldsymbol{S}_k$, we obtain $p(X_m|\boldsymbol{S})=p(X_m|\boldsymbol{S}_k)$. Thus the estimation of $p(X_m|\boldsymbol{S})$ is biased towards the conditional features $\boldsymbol{S}_k$ used. In Section 2.2, we will show how some existing methods attempt to reduce this bias.

2.2 Methods Based on Geometric Mean Distribution

First let us describe the concept of k-combination. A k-combination of a set $\boldsymbol{S}$ is a subset of k distinct elements of $\boldsymbol{S}$. The number of k-combinations of the set $\boldsymbol{S}$ is equal to the binomial coefficient $\binom{|\boldsymbol{S}|}{k}=\frac{|\boldsymbol{S}|!}{k!(|\boldsymbol{S}|-k)!}$. We denote the i_{th} k-combination of $\boldsymbol{S}$ as $\boldsymbol{S}_k^i$ and all the possible k-combinations of $\boldsymbol{S}$ as $\mathbb{S}_k$. Thus $\boldsymbol{S}_k^i\in\mathbb{S}_k$, where $1\leq i\leq\binom{|\boldsymbol{S}|}{k}$. The *geometric mean distribution* assumption is then defined as:

Geometric Mean Distribution (GMD). The (class-conditional) probability density function of a candidate feature X_m given the selected features $\boldsymbol{S}$ is equal to the geometric mean of the (class-conditional) probability density function of X_m conditioning on any k ($0\leq k\leq|\boldsymbol{S}|$) features in $\boldsymbol{S}$:

$$p(X_m|\boldsymbol{S})\simeq\Big(\prod_{\boldsymbol{S}_k^i\in\mathbb{S}_k} p(X_m|\boldsymbol{S}_k^i)\Big)^{\frac{1}{\binom{|\boldsymbol{S}|}{k}}}, \tag{7}$$

$$p(X_m|\boldsymbol{S},C)\simeq\Big(\prod_{\boldsymbol{S}_k^i\in\mathbb{S}_k} p(X_m|\boldsymbol{S}_k^i,C)\Big)^{\frac{1}{\binom{|\boldsymbol{S}|}{k}}}. \tag{8}$$

This GMD assumption is a relaxed version of FID in the sense that if FID holds, GMD is true. Using the geometric mean of probability densities across all conditional feature subsets $\mathbb{S}_k$, GMD assumption removes the estimation bias towards any specific conditional feature subsets $\boldsymbol{S}_k$. As before, k is a hyper-parameter that controls the information loss and model complexity. Note that Eq. (8) with $k=1$ has been used in (Gao, Ver Steeg, and Galstyan 2016).

Theorem 2. *Under the GMD assumption of order k, the objective function in Eq. (2) becomes:*

$$J_{\mathrm{GMD}}^{\mathrm{k,k}}(X_m)\sim\frac{1}{\binom{|\boldsymbol{S}|}{k}}\sum_{\boldsymbol{S}_k^i\in\mathbb{S}_k} I(X_m;C|\boldsymbol{S}_k^i). \tag{9}$$

More generally, we consider different values of k in Eq. (7) and (8), denoted as k_1 and k_2 respectively. The objective function $J(X_m)$ is equivalent to:

$$\begin{aligned} J_{\mathrm{GMD}}^{\mathrm{k_1,k_2}}(X_m)\sim&\frac{1}{\binom{|\boldsymbol{S}|}{k_1}}\sum_{\boldsymbol{S}_{k_1}^i\in\mathbb{S}_{k_1}} H(X_m|\boldsymbol{S}_{k_1}^i)\\ &-\frac{1}{\binom{|\boldsymbol{S}|}{k_2}}\sum_{\boldsymbol{S}_{k_2}^i\in\mathbb{S}_{k_2}} H(X_m|\boldsymbol{S}_{k_2}^i,C). \end{aligned} \tag{10}$$

In Theorem 2, we have derived a class of methods that use the GMD assumption, and their objective function is shown in Eq. (10). See Appendix I.B for the proof.

Existing Methods Using GMD. By varying k_1 and k_2 in Theorem 2, we recover four existing MI based methods, i.e., MIM, Minimum-Redundancy Maximum-Relevance (MRMR) (Peng, Long, and Ding 2005), Joint Mutual Information (JMI) (Yang and Moody 1999; Meyer, Schretter, and Bontempi 2008) and Relaxed Minimum-Redundancy Maximum-Relevance (RMRMR) (Vinh et al. 2016). Their

objective functions are listed in Table 1 (J_{GMD}), and the proof is shown in Appendix I.B. Note that GMD has not been explicitly used to derive the objective functions for these methods before. *We contribute here by identifying the GMD assumption as their common theoretical foundations and showing these methods differ only in order k.*

Time Complexity of GMD Methods. To analyze the time complexity of the methods presented in Theorem 2 with different k_1 and k_2 values ($J^{\mathrm{k_1,k_2}}_{\mathrm{GMD}}$), we assume all the joint entropy or MI values can be calculated in $\mathcal{O}(N)$ time. Let $t = \max\{k_1, k_2\}$, and K denote the maximum number of features to be selected. The number of entropy values needs to be calculated is in $\mathcal{O}\big(KM\binom{K}{t-1}\big)$. Thus, the overall time complexity is $\mathcal{O}\big(KMN\binom{K}{t-1}\big)$ if $t \geq 1$, and $\mathcal{O}(MN)$ if $t = 0$.

Strength and Limitation of GMD. The GMD assumption is tractable and reasonable as a relaxation of the FID assumption. The probability density estimates under the GMD assumption are not biased towards the conditional features used. However, the GMD assumption is more complex and it is hard to describe the exact condition under which the GMD assumption holds. Furthermore, these probability density estimates may not sum up to 1, because the geometric mean of nonnegative real numbers is less than or equal to their arithmetic mean, which sums up to 1:

$$\begin{aligned}\sum_{X_m} p(X_m|\boldsymbol{S}) &\simeq \sum_{X_m}\Big(\prod_{\boldsymbol{S}^i_k\in\mathbb{S}_k} p(X_m|\boldsymbol{S}^i_k)\Big)^{\frac{1}{\binom{|\boldsymbol{S}|}{k}}} \\ &\leq \sum_{X_m}\frac{1}{\binom{|\boldsymbol{S}|}{k}}\Big(\sum_{\boldsymbol{S}^i_k\in\mathbb{S}_k} p(X_m|\boldsymbol{S}^i_k)\Big) = 1.\end{aligned} \tag{11}$$

where the equality holds only when all $p(X_m|\boldsymbol{S}^i_k)$ are equal. The unnormalization of probability densities may have a large impact on the calculation of MI, e.g., resulting in a negative MI value. One possible way to solve this issue is to divide probability densities by a normalization constant (Wolpert and Wolf 1995; Rajwade, Banerjee, and Rangarajan 2008; Berglund, Raiko, and Cho 2015). However, computing the normalization constant is nontrivial in our case.

2.3 Methods Based on Arithmetic Mean Distribution

In this subsection, we introduce a logical extension of the GMD assumption that uses arithmetic mean to estimate probability densities:

Arithmetic Mean Distribution (AMD). The (class-conditional) probability density function of a candidate feature X_m given the selected features $\boldsymbol{S}$ is equal to the arithmetic mean of the (class-conditional) probability density function of X_m conditioning on any k ($0 \leq k \leq |\boldsymbol{S}|$) features in $\boldsymbol{S}$:

$$p(X_m|\boldsymbol{S}) \simeq \frac{1}{\binom{|\boldsymbol{S}|}{k}}\Big(\sum_{\boldsymbol{S}^i_k\in\mathbb{S}_k} p(X_m|\boldsymbol{S}^i_k)\Big), \tag{12}$$

$$p(X_m|\boldsymbol{S}, C) \simeq \frac{1}{\binom{|\boldsymbol{S}|}{k}}\Big(\sum_{\boldsymbol{S}^i_k\in\mathbb{S}_k} p(X_m|\boldsymbol{S}^i_k, C)\Big). \tag{13}$$

The AMD assumption is another relaxation of the FID assumption. But unlike GMD, the probability densities estimated by AMD are always normalized, i.e., sum up to 1.

Methods Using AMD. Consider the objective function:

$$I(X_m; C|\boldsymbol{S}) = \sum_{X_m,\boldsymbol{S},C} p(X_m, \boldsymbol{S}, C)\log\frac{p(X_m|\boldsymbol{S}, C)}{p(X_m|\boldsymbol{S})}. \tag{14}$$

We use sample mean to approximate the expectation term

$$\hat{I}(X_m; C|\boldsymbol{S}) = \frac{1}{N}\sum_{X^{(i)}_m,\boldsymbol{S}^{(i)},C^{(i)}} \log\frac{p\big(X^{(i)}_m|\boldsymbol{S}^{(i)}, C^{(i)}\big)}{p\big(X^{(i)}_m|\boldsymbol{S}^{(i)}\big)}, \tag{15}$$

where $\{X^{(i)}_m, \boldsymbol{S}^{(i)}, C^{(i)}\}$ denotes the i_{th} sample, N is sample size, and $\hat{I}(\cdot)$ denotes the sample estimate for $I(\cdot)$. By substituting $p(X_m|\boldsymbol{S})$ and $p(X_m|\boldsymbol{S}, C)$ using AMD, we obtain a ***new*** class of methods, denoted as $J^{\mathrm{k_1,k_2}}_{\mathrm{AMD}}$, where k_1 and k_2 are the values of k in Eq. (12) and (13) respectively. We will consider four pairs of k_1 and k_2 values in our experiments, $J^{0,0}_{\mathrm{AMD}}$, $J^{1,0}_{\mathrm{AMD}}$, $J^{1,1}_{\mathrm{AMD}}$, and $J^{2,1}_{\mathrm{AMD}}$, to investigate the effects of order k. Note that $J^{0,0}_{\mathrm{AMD}}$ is essentially equivalent to $J^{0,0}_{\mathrm{GMD}}$ and $J^{0,0}_{\mathrm{FID}}$, that are all the same as the MIM method.

Time Complexity of AMD Methods. Although the high-dimensional MI, i.e., $I(X_m; C|\boldsymbol{S})$, cannot be decomposed into a series of low-order MI approximations when using the AMD assumption, the calculation of $I(X_m; C|\boldsymbol{S})$ is still tractable, simply because AMD is tractable to compute. We note the time complexity of the AMD methods ($J^{\mathrm{k_1,k_2}}_{\mathrm{AMD}}$) is the same as that of the methods based on GMD ($J^{\mathrm{k_1,k_2}}_{\mathrm{GMD}}$).

Strength and Limitation of AMD. The AMD assumption is a reasonable relaxation of the FID assumption, in the sense that if FID holds, AMD is true. The probability density estimates under the AMD assumption are unbiased and always sum up to one. But like GMD, it is hard to describe the exact condition under which AMD holds. Furthermore, the high-dimensional MI cannot be simply written as a series of low-order MI approximations under the AMD assumption.

3 Variational Information Based Methods

Instead of estimating high-order MI using low-order approximations, a lower bound on MI called Variational Information (VI) can be used to design objective function for feature selection (Gao, Ver Steeg, and Galstyan 2016).

3.1 Variational Mutual Information

Consider two random variables X and Y and their joint probability distribution $p(X, Y)$. A lower bound on the MI between X and Y is defined as (Barber and Agakov 2004):

$$I(X;Y) \geq H(Y) + \sum_{X,Y} p(X,Y)\log q(Y|X) := I_{\mathrm{lb}}(X;Y), \tag{16}$$

where $q(Y|X)$ is an arbitrary variational distribution as long as it is normalized. This lower bound is derived based on the Kullback-Leibler divergence i.e., $\sum_Y p(Y|X)\log p(Y|X) - p(Y|X)\log q(Y|X) \geq 0$, and becomes exact when the variational distribution $q(Y|X)$ matches the real one $p(Y|X)$ (Barber and Agakov 2004).

3.2 Methods Based on Variational Information

The VI bound has been used to design objective function for feature selection in (Gao, Ver Steeg, and Galstyan 2016):

$$I_{\mathrm{lb}}(\boldsymbol{T}; C) := \sum_{\boldsymbol{T}, C} p(\boldsymbol{T}, C) \log \frac{q(\boldsymbol{T}|C)}{\sum\limits_{C} q(\boldsymbol{T}|C) p(C)}, \quad (17)$$

where $\boldsymbol{T} = X_m \cup \boldsymbol{S}$ defines a trial feature subset. It has been verified that this lower bound is valid for any distribution $q(\boldsymbol{T}|C)$ used in Eq. (17), and becomes exact if $q(C|\boldsymbol{T}) = p(C|\boldsymbol{T})$ (Gao, Ver Steeg, and Galstyan 2016).

For Sequential Forward Selection, maximizing $I_{\mathrm{lb}}(\boldsymbol{T}; C)$ is equivalent to maximizing $I_{\mathrm{lb}}(X_m; C|\boldsymbol{S})$:

$$I_{\mathrm{lb}}(X_m; C|\boldsymbol{S}) := \sum_{X_m, \boldsymbol{S}, C} p(X_m, \boldsymbol{S}, C) \log \frac{q(X_m|\boldsymbol{S}, C)}{\sum\limits_{C} q(X_m|\boldsymbol{S}, C) p(C)}. \quad (18)$$

Using sample mean to estimate the expectation term yields

$$\tilde{J}(X_m) \sim \frac{1}{N} \sum_{X_m^{(i)}, \boldsymbol{S}^{(i)}, C^{(i)}} \log \frac{q\big(X^{(i)}|\boldsymbol{S}^{(i)}, C^{(i)}\big)}{\sum\limits_{C} q\big(X^{(i)}|\boldsymbol{S}^{(i)}, C^{(i)}\big) p\big(C^{(i)}\big)}, \quad (19)$$

where $\{X_m^{(i)}, \boldsymbol{S}^{(i)}, C^{(i)}\}$ denotes the i_{th} sample, N is sample size. We denote the objective function of VI methods as $\tilde{J}(\cdot)$, in contrast to that of MI methods $J(\cdot)$.

In order to calculate the lower bound, we need to specify a proper choice of probability distribution $q(X_m|\boldsymbol{S}, C)$ that is both realistic and tractable to compute. Note that the VI based methods only require the class-conditional probability distribution of features (i.e., $q(X_m|\boldsymbol{S}, C)$), in contrast to the MI based methods which require both conditional and class-conditional probability distributions of features (i.e., $p(X_m|\boldsymbol{S})$ and $p(X_m|\boldsymbol{S}, C)$). In this sense, the VI based methods are more relaxed than the MI based methods.

The FID, GMD and AMD assumptions can also be plugged in Eq. (18) to compute $q(X_m|\boldsymbol{S}, C)$ and thus the objective function for the VI based methods. In fact the GMD assumption, Eq. (8) with $k = 0$ and $k = 1$, has been tested in (Gao, Ver Steeg, and Galstyan 2016). As discussed before, the GMD assumption may result in an unnormalized estimate of probability densities. In the experiments section, we will investigate whether the AMD assumption can lead to the selection of more informative features by fixing the normalization issue of GMD. We note that the time complexity of the VI based methods is the same as the corresponding MI based methods using the same distribution assumption.

4 Experiments

In this section, we systematically compare the performance of three probability distribution assumptions, i.e., FID, GMD and AMD. We combine each assumption with the MI and VI frameworks, resulting in a number of different feature selection methods. *The Python and C++ source codes of these methods are available in the supplementary material*. These methods are then used to select a subset of features for 29 real-world classification tasks (see Table 2 for details[1]). The features selected by each method are evaluated using two classifiers – K Nearest Neighbour (KNN) with $K = 3$ and linear Support Vector Machine (SVM) with the regularization parameter set to 1. We calculate the average 10-folder cross-validation error rate on the range of 10 to 100 features (or 10 to M if the number of features $M < 100$) as an indication of the effectiveness of feature selection methods, following (Nguyen et al. 2014; Gao, Ver Steeg, and Galstyan 2016). This process is repeated 50 times to alliterative randomness. The Wilcoxon rank-sum test ($\alpha = 0.05$) with Holm p-value correction (Sheskin 2003) is used to determine statistical significance. For datasets with continuous features, the Minimum Description Length method (Fayyad and Irani 1993) is employed to evenly divide the continuous values into five bins, following (Vinh et al. 2016). Note that the discretization is only used in the feature selection procedure, while the classifiers still use the original continuous values.

Table 2: Description of the data sets used in the experiments. M: the number of features; N: the number of samples; C: the number of classes.

ID	Name	M	N	C	M/N	MN
d_1	Allaml	7129	72	2	9.9e+01	5.1e+05
d_2	Breast	30	569	2	5.2e-02	1.7e+04
d_3	Carcinom	9182	174	11	5.2e+01	1.6e+06
d_4	Coil20	1024	1440	20	7.1e-01	1.4e+06
d_5	Colon	2000	62	2	3.2e+01	1.2e+05
d_6	Congress	16	435	2	3.6e-02	6.9e+03
d_7	Glioma	4434	50	4	8.8e+01	2.2e+05
d_8	Heart	13	270	2	4.8e-02	3.5e+03
d_9	Ionosphere	34	351	2	9.6e-02	1.1e+04
d_{10}	Isolet	617	1560	26	3.9e-01	9.6e+05
d_{11}	Krvskp	36	3196	2	1.1e-02	1.1e+05
d_{12}	Landsat	36	6435	6	5.5e-03	2.3e+05
d_{13}	Leuk	7070	72	2	9.8e+01	5.0e+05
d_{14}	Lung	3312	203	5	1.6e+01	6.7e+05
d_{15}	Lungcancer	56	32	3	1.7e+00	1.7e+03
d_{16}	Lymphoma	4026	96	9	4.1e+01	3.8e+05
d_{17}	NCI9	9712	60	9	1.6e+02	5.8e+05
d_{18}	ORL	1024	400	40	2.5e+00	4.1e+05
d_{19}	Parkinsons	22	195	2	1.1e-01	4.2e+03
d_{20}	Semeion	256	1593	10	1.6e-01	4.0e+05
d_{21}	Sonar	60	208	2	2.8e-01	1.2e+04
d_{22}	Soybeansmall	35	47	4	7.4e-01	1.6e+03
d_{23}	Spect	22	267	2	8.2e-02	5.8e+03
d_{24}	Splice	60	3175	3	1.8e-02	1.9e+05
d_{25}	TOX_171	5748	171	4	3.3e+01	9.8e+05
d_{26}	WarpAR10P	2400	130	10	1.8e+01	3.1e+05
d_{27}	WarpPIE10P	2420	210	10	1.1e+01	5.0e+05
d_{28}	Waveform	40	5000	3	8.0e-03	2.0e+05
d_{29}	Wine	13	178	3	7.3e-02	2.3e+03

[1]The datasets are available from the UCI Machine Learning Repository (Lichman 2013) http://archive.ics.uci.edu/ml/.

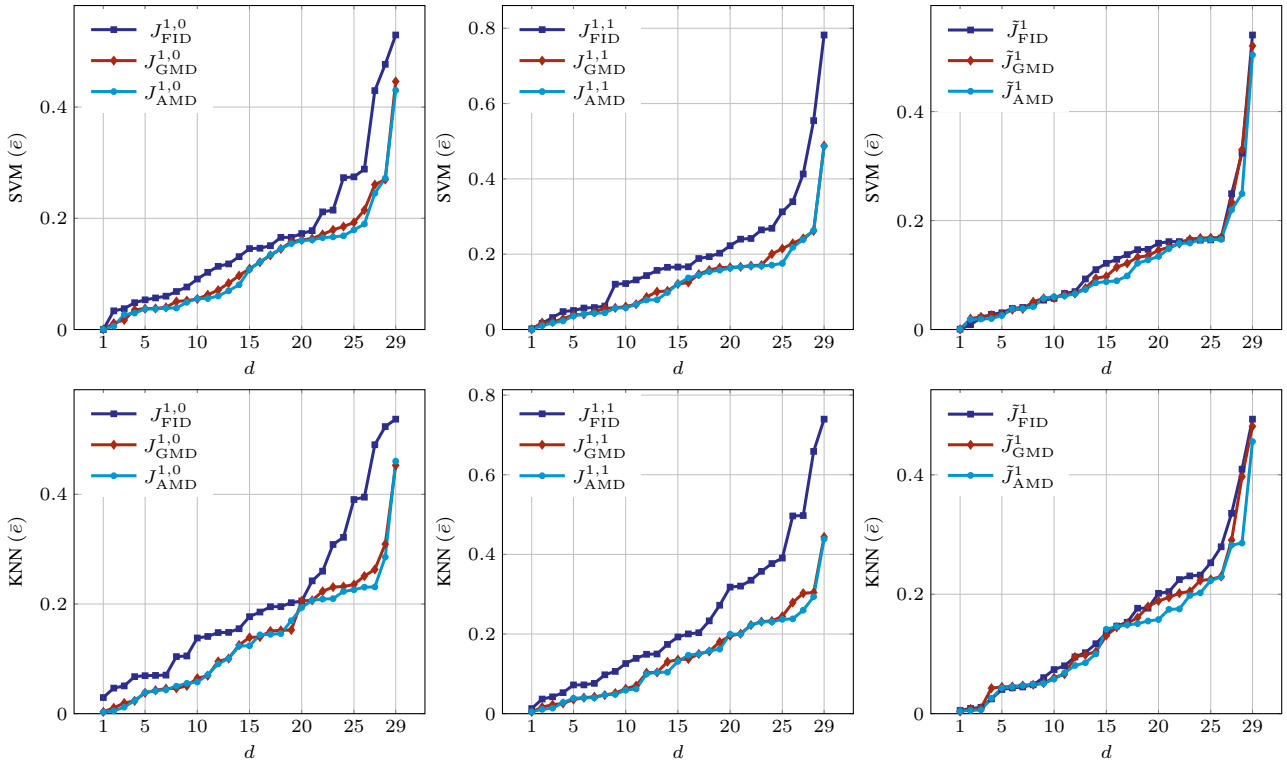

Figure 1: Comparison between the FID, GMD and AMD assumptions incorporated with the MI (left and middle figures) and VI frameworks (right figures). The horizontal axis represents the datasets (d), and the vertical axis the mean classification error rates ($\bar{e}$) generated by KNN or SVM using the features selected by each method. For each method, we sort the error rates on the 29 datasets in ascending order to generate the plots for better visualization.

4.1 Comparison of Probability Distributions Within the MI and VI Frameworks

Setup We compare the efficacy of FID, GMD and AMD assumptions, each within the MI and VI frameworks. We denote the objective function of MI based methods using J, while the VI based methods as $\tilde{J}$. To see the effects of different probability distribution assumptions, we fix the order k (i.e., the number of conditional features), and only vary the assumptions. Specifically, we consider two sets of comparisons for the MI based methods: 1) $J^{1,0}_{\text{AMD}}$ vs. $J^{1,0}_{\text{GMD}}$ (MRMR) vs. $J^{1,0}_{\text{FID}}$ (MIFS with $\beta = 1$); and 2) $J^{1,1}_{\text{AMD}}$ vs. $J^{1,1}_{\text{GMD}}$ (JMI) vs. $J^{1,1}_{\text{FID}}$ (CIFE). Note, unlike MI based methods, the VI based methods only use the class-conditional feature distributions, thus we only consider one set of comparison for VI based methods: $\tilde{J}^{1}_{\text{AMD}}$ vs. $\tilde{J}^{1}_{\text{FID}}$ vs. $\tilde{J}^{1}_{\text{GMD}}$ (i.e., VMI_{pairwise} (Gao, Ver Steeg, and Galstyan 2016)). We do not compare $\tilde{J}^{0}_{\text{FID}}$ vs. $\tilde{J}^{0}_{\text{GMD}}$ vs. $\tilde{J}^{0}_{\text{AMD}}$ because they are equivalent in the sense that FID, GMD and AMD degenerate to the same distribution when $k = 0$.

Results The classification error rates of KNN and SVM using the features selected by these methods under different probability distribution assumptions are plotted in Figure 1. To generate these plots, we sort the classification error rates produced by each method on the 29 datasets in ascending order. Therefore, the dataset index in these figures does not match that in Table 2. We also rank these methods according to their classification accuracy on each dataset using statistical tests. The average ranking of each method is shown in Table 3. The detailed classification error rates on each dataset can be found in Appendix Table A1 and A2.

Table 3: The average ranking of the classification error rates generated by the methods based on the FID, GMD and AMD assumptions. The **best** ranking is highlighted in bold.

Methods	Classifier	FID	GMD	AMD
$J^{1,0}$	SVM	2.52	1.79	**1.31**
	KNN	2.62	1.59	**1.44**
$J^{1,1}$	SVM	2.27	1.90	**1.24**
	KNN	2.69	1.69	**1.24**
$\tilde{J}^{1}$	SVM	2.17	1.83	**1.38**
	KNN	2.28	1.93	**1.38**

We observe that the feature selection methods based on AMD and GMD generally achieve much lower classification error than those based on FID. It confirms our hypothesis that the AMD and GMD assumptions can potentially reduce the estimation bias of FID. Moreover, the methods based on AMD yield equally well or statistically significantly better

Table 4: The average ranking of the classification accuracy generated by the methods based on GMD and AMD with different order k. The **best** ranking is highlighted in bold.

Distribution	Classifier	$J^{0,0}$	$J^{1,0}$	$J^{1,1}$	$J^{2,1}$
GMD	SVM	3.34	1.59	2.41	**1.48**
	KNN	3.41	2.03	2.20	**1.55**
AMD	SVM	3.31	1.83	2.38	**1.79**
	KNN	3.18	2.14	2.20	**1.82**

classification accuracy than those based on GMD, and the average ranking of AMD-based methods is consistently better than those based on GMD. It empirically demonstrates that by fixing the normalization issue of GMD using AMD, we can potentially select more informative features, resulting in a higher classification accuracy.

4.2 Effects of Order k in Probability Distributions

Setup To investigate the effects of order k on the performance of feature selection methods, we fix the probability distribution assumptions and only vary k. Specifically, we consider two sets of comparisons: 1) $J^{0,0}_{\text{AMD}}$ (MIM) vs. $J^{1,0}_{\text{AMD}}$ vs. $J^{1,1}_{\text{AMD}}$ vs. $J^{2,1}_{\text{AMD}}$; and 2) $J^{0,0}_{\text{GMD}}$ (MIM) vs. $J^{1,0}_{\text{GMD}}$ (MRMR) vs. $J^{1,1}_{\text{GMD}}$ (JMI) vs. $J^{2,1}_{\text{GMD}}$ (RMRMR). We limit the value of k up to 2, because: 1) the existing methods based on GMD (J_{GMD}) are mainly low-order, and 2) estimating a higher-order probability distribution requires significantly more samples. We do not consider the VI-based methods here because, when $k = 2$, the VI based methods require to compute probability distributions conditional on three variables: two features plus one class label.

Results For each dataset, we rank the methods with different order k based on their SVM or KNN classification error rates. The average ranking of each method is shown in Table 4, and the detailed classification error rates can be found in Appendix Table A1 and A2. The plots of classification error rates and convergence curves are placed in Appendix Figure A2 and A3 due to page limits. We observe that the performance of these methods based on the GMD and AMD distributions improves as k increases from 0 to 2. This is not surprising because, in theory as k increases, the GMD and AMD assumptions become more realistic. However, increasing the value of k also significantly increases the computational cost, as shown in Appendix Figure A4.

4.3 Comparison Between the Methods Based on AMD Against Other State-of-the-arts

Setup In Appendix III.C, we have found that the two methods based on AMD, $\tilde{J}^{1}_{\text{AMD}}$ and $J^{2,1}_{\text{AMD}}$, are very competitive among the 12 methods considered. Here we further compare $\tilde{J}^{1}_{\text{AMD}}$ and $J^{2,1}_{\text{AMD}}$ against three other information-theoretic methods that have not been included in our framework – MRI (Wang et al. 2017), $\mathcal{SPEC}_{\text{CMI}}$ (Nguyen et al. 2014) and CMIM (Fleuret 2004), as well as two methods that are not based on information theory – Trace (Nie et al. 2008), and SPEC (Zhao and Liu 2007).

Table 5: The average ranking of the classification accuracy generated by the methods based on AMD compared to other state-of-the-arts. The **best** and *second best* ranking are highlighted in bold and italics respectively.

Classifier	$\tilde{J}^{1}_{\text{AMD}}$	$J^{2,1}_{\text{AMD}}$	MRI	$\mathcal{SPEC}_{\text{CMI}}$	CMIM	Trace	SPEC
SVM	*2.58*	**2.34**	2.86	4.24	2.72	4.82	6.52
KNN	**2.41**	*2.65*	3.00	4.17	3.03	4.62	6.31

Results The average ranking results in Table 5 show that the two methods based on the AMD assumption achieve overall better performance than the other state-of-the-art methods. It confirms that the AMD assumption on feature distributions is reasonable and can be used to select informative features for many real-world datasets. The detailed classification error rates of each method on each dataset are presented in Appendix Table A3 and A4.

5 Conclusion

In this paper, we have revealed the key role of the *Feature Independence Distribution* (FID) and *Geometric Mean Distribution* (GMD) assumptions used in many information-theoretic feature selection methods. We showed that the probability density estimates under the FID assumption are biased, and the GMD assumption, although reducing this estimation bias, introduces an additional normalization issue, i.e., its probability density estimates may not sum up to one. We resolved this issue by proposing the *Arithmetic Mean Distribution* (AMD) assumption. Our numerical experiments confirmed that the AMD assumption improves over the GMD and FID assumptions in selecting informative features within both the Mutual Information and Variational Information frameworks.

Acknowledgments

This work was partially supported by an ARC Discovery Grant (DP180101170) from Australian Research Council.

References

Bagherzadeh-Khiabani, F.; Ramezankhani, A.; Azizi, F.; Hadaegh, F.; Steyerberg, E. W.; and Khalili, D. 2016. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. *Journal of Clinical Epidemiology* 71:76–85.

Balagani, K. S., and Phoha, V. V. 2010. On the feature selection criterion based on an approximation of multidimensional mutual information. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (7):1342–1343.

Barber, D., and Agakov, F. 2004. The IM algorithm: a variational approach to information maximization. *Advances in Neural Information Processing Systems* 16:201.

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

Berglund, M.; Raiko, T.; and Cho, K. 2015. Measuring the usefulness of hidden units in boltzmann machines with mutual information. *Neural Networks* 64:12–18.

Bolón-Canedo, V.; Sánchez-Maroño, N.; and Alonso-Betanzos, A. 2015. Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-Based Systems* 86:33–45.

Brown, G.; Pocock, A.; Zhao, M.-J.; and Luján, M. 2012. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of Machine Learning Research* 13(Jan):27–66.

Cai, J.; Luo, J.; Wang, S.; and Yang, S. 2018. Feature selection in machine learning: A new perspective. *Neurocomputing* 300:70–79.

Cover, T. M., and Thomas, J. A. 2012. *Elements of information theory*. John Wiley & Sons.

Fayyad, U., and Irani, K. 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, 1022–1029.

Fleuret, F. 2004. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research* 5(Nov):1531–1555.

Gao, S.; Ver Steeg, G.; and Galstyan, A. 2016. Variational information maximization for feature selection. In *Advances in Neural Information Processing Systems*, 487–495.

Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3(Mar):1157–1182.

Kittler, J. 1986. Feature selection and extraction. *Handbook of Pattern Recognition and Image Processing* 59–83.

Lewis, D. D. 1992. Feature selection and feature extraction for text categorization. In *Proceedings of the Workshop on Speech and Natural Language*, 212–217. Association for Computational Linguistics.

Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; and Liu, H. 2017. Feature selection: A data perspective. *ACM Computing Surveys* 50(6):94.

Lichman, M. 2013. UCI machine learning repository. *University of California, Irvine, School of Information and Computer Sciences*.

Lin, D., and Tang, X. 2006. Conditional infomax learning: an integrated framework for feature extraction and fusion. In *European Conference on Computer Vision*, 68–82. Springer.

Meyer, P. E.; Schretter, C.; and Bontempi, G. 2008. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing* 2(3):261–274.

Nguyen, X. V.; Chan, J.; Romano, S.; and Bailey, J. 2014. Effective global approaches for mutual information based feature selection. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 512–521. ACM.

Nie, F.; Xiang, S.; Jia, Y.; Zhang, C.; and Yan, S. 2008. Trace ratio criterion for feature selection. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, volume 2, 671–676.

Noshad, M.; Zeng, Y.; and Hero, A. O. 2019. Scalable mutual information estimation using dependence graphs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2962–2966. IEEE.

Pascoal, C.; De Oliveira, M. R.; Valadas, R.; Filzmoser, P.; Salvador, P.; and Pacheco, A. 2012. Robust feature selection and robust pca for internet traffic anomaly detection. In *2012 Proceedings IEEE INFOCOM*, 1755–1763. IEEE.

Peng, H.; Long, F.; and Ding, C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(8):1226–1238.

Pudil, P.; Novovičova, J.; and Kittler, J. 1994. Floating search methods in feature selection. *Pattern Recognition Letters* 15(11):1119–1125.

Rajwade, A.; Banerjee, A.; and Rangarajan, A. 2008. Probability density estimation using isocontours and isosurfaces: applications to information-theoretic image registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(3):475–491.

Rodriguez-Lujan, I.; Huerta, R.; Elkan, C.; and Cruz, C. S. 2010. Quadratic programming feature selection. *Journal of Machine Learning Research* 11(Apr):1491–1516.

Saeys, Y.; Inza, I.; and Larrañaga, P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19):2507–2517.

Sharmin, S.; Shoyaib, M.; Ali, A. A.; Khan, M. A. H.; and Chae, O. 2019. Simultaneous feature selection and discretization based on mutual information. *Pattern Recognition* 91:162–174.

Sheskin, D. J. 2003. *Handbook of parametric and nonparametric statistical procedures*. crc Press.

Singha, S., and Shenoy, P. P. 2018. An adaptive heuristic for feature selection based on complementarity. *Machine Learning* 107(12):2027–2071.

Venkateswara, H.; Lade, P.; Lin, B.; Ye, J.; and Panchanathan, S. 2015. Efficient approximate solutions to mutual information based global feature selection. In *2015 IEEE International Conference on Data Mining*, 1009–1014. IEEE.

Vergara, J. R., and Estévez, P. A. 2014. A review of feature selection methods based on mutual information. *Neural Computing and Applications* 24(1):175–186.

Vinh, N. X.; Zhou, S.; Chan, J.; and Bailey, J. 2016. Can high-order dependencies improve mutual information based feature selection? *Pattern Recognition* 53:46–58.

Wang, J.; Wei, J.-M.; Yang, Z.; and Wang, S.-Q. 2017. Feature selection by maximizing independent classification information. *IEEE Transactions on Knowledge and Data Engineering* 29(4):828–841.

Wolpert, D. H., and Wolf, D. R. 1995. Estimating functions of probability distributions from a finite set of samples. *Physical Review E* 52(6):6841.

Wu, Y., and Yang, P. 2016. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory* 62(6):3702–3720.

Xue, B.; Zhang, M.; Browne, W. N.; and Yao, X. 2016. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation* 20(4):606–626.

Yang, H. H., and Moody, J. 1999. Data visualization and feature selection: New algorithms for nongaussian data. In *Advances in Neural Information Processing Systems*, 687–693.

Zadeh, S. A.; Ghadiri, M.; Mirrokni, V.; and Zadimoghaddam, M. 2017. Scalable feature selection via distributed diversity maximization. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Zhao, Z., and Liu, H. 2007. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th International Conference on Machine Learning*, 1151–1157. ACM.