Information Sufficiency via Fourier Expansion

Mohsen Heidari
Purdue University
mheidari@purdue.edu

Jithin Sreeharam
CSoI
jithin.k.s@gmail.com

Gil Shamir
Google Inc.
gshamir@google.com

Wojciech Szpankowski
Purdue University
szpan@purdue.edu

***Abstract*—We take an information-theoretic approach to identify nonlinear feature redundancies in unsupervised learning. We define a subset of features as sufficiently-informative when the joint entropy of all the input features equals to that of the chosen subset. We argue that the rest of the features are redundant as all the accessible information about the data can be captured from sufficiently-informative features. Next, instead of directly estimating the entropy, we propose a Fourier-based characterization. For that, we develop a novel Fourier expansion on the Boolean cube incorporating correlated random variables. This generalization of the standard Fourier analysis is beyond product probability spaces. Based on our Fourier framework, we propose an algorithm for unsupervised feature selection. Via a theoretical analysis, we show that our proposed algorithms find provably *asymptotically optimal* feature subsets. Also, numerical experiments demonstrate that our methods outperform state-of-the-art feature selection algorithms on various datasets. Lastly, we show that the computational complexity of our algorithms can be as low as $O(nd)$ with n being the number of samples and d the number of features.**

I. Introduction

A central challenge in learning with feature selection is to jointly identify nonlinear *redundancies* within the features and the *dependencies* in the feature-label relation. Many well-known feature selection approaches (supervised or unsupervised) are based on measures that capture only linear relations or focus on the features individually [1]–[3]. Kernel-based methods are exception; however, are prohibitive in large datasets as the computational complexity of computing a kernel grows super linearly with the number of the samples [4]. Alternatively, information-theoretic metrics are powerful candidates in quantifying nonlinear dependencies among the random variables. However, estimating such quantities usually requires high sample complexity.

In this work, we take an alternative approach and adapt discrete Fourier analysis with information theoretic measures. Hence, capturing nonlinear relations with low sample complexity, while avoiding kernel computations. The discrete Fourier expansion (on the Boolean cube) provides an essential tool to characterize different levels of "nonlinearities" in a function. In this expansion, any real-valued function on the Boolean cube can be written as a linear combination of monomials (*parities*) [5], [6]. Highly nonlinear functions have Fourier expansion with large coefficients for high-degree parities. Thus, the Fourier expansion is potentially a powerful tool in learning problems. However, there are limitations making it impractical. First, it is developed for product probability spaces (mutually independent input variables). Secondly, this expansion is defined only for deterministic functions. These assumptions are too strong, as learning problems, often, involve correlated features with stochastic labeling. In this work, we make a connection between the two approaches and aim to address these challenges. A full version of the paper cab be found in [7].

A. Main Contributions

We propose entropy as a justifiable measure of feature redundancy and identify the notion of sufficiently informative for feature subsets. We address the above challenges by developing a novel Fourier expansion for stochastic mappings of *correlated* binary random variables. Using this framework, we study feature selection, where the objective is to remove as many features as possible without significantly increasing classification loss. We demonstrate that the Fourier expansion provides a powerful tool to characterize nonlinear *redundancies* in features and nonlinear *dependencies* in the features-label relation. We propose a Fourier-based algorithm for unsupervised variant of the problem. We use the Fourier expansion to provide a theoretical analysis and derive conditions under which our algorithm finds the optimal feature subset. Further, through numerical experiments, we show that our algorithm outperforms several well-known feature selection techniques. That said, the contributions of this paper are three-fold as summarized below:

1) Fourier expansion for correlated random variables: We develop a generalized Fourier expansion for functions of *correlated binary* random variables (Proposition 1). For this purpose, we adopt a Gram-Schmidt-type orthogonalization and construct a set of orthogonal basis functions. Further, we adapt our Fourier expansion to the more general space of stochastic mappings (e.g., mappings from one probability space to another). To the best of our knowledge, this is the first generalization of the Fourier expansion for correlated binary random variables. Although this Fourier expansion is defined on the Boolean cube, our algorithms are applicable to non-binary features too. We view the Binary Fourier as a framework that captures a special class of nonlinearities — those characterized via the *parities*. Alternatively, we could generalize our Fourier expansion to discrete features and, based on it, design feature selection algorithms. However, such a generalization requires character theory, which is beyond the scope of this paper. We note that there are other forms of orthogonal decomposition including the Hoeffding-Sobel decomposition [8]–[10] and its generalization [10]. However,

such decompositions are basis-free. Our Fourier expansion is defined by constructing a set of *orthonormal* basis functions which makes it suitable for feature selection.

2) Sufficiently Informative: In the unsupervised setting, we take an information-theoretic perspective, and group the features into *redundant* and *sufficiently informative*. All the accessible *information* about the data can be captured from the later group. More precisely, we define a subset of features as sufficiently-informative when the joint entropy of all the input features equals to that of the chosen subset. The former group is statistically a function of the later, hence can be removed without affecting the learning's performance. This approach extends the notion of *Markov blanket* for "redundant" features to the unsupervised setting [11]–[13]. We then develop a characterization of sufficiently informative features based on our Fourier expansion (Theorem 1). Built upon this, we design an Unsupervised Fourier Feature Selection (UFFS) algorithm, which captures the redundant features in our new formulation. Instead of ranking the features, the UFFS finds redundant features and declares the rest of the features as sufficiently informative. Through comprehensive numerical experiments in Section V, we show that the UFFS significantly outperforms well-known methods for unsupervised feature selection.

B. Related Works

Our unsupervised feature selection algorithm is multi-variate (evaluating the dependencies of features jointly) and identifies non-linear dependencies. Whereas many well-known existing works are considered univariate or focus on the linear dependencies among the features [3]. Some common approaches in unsupervised feature selection are pseudo-label based, "column subset selection", and spectral/manifold based. Methods in the first approach such as [14], [15] attempt to generate pseudo-labels via certain clustering methods. However, such methods focus on linear transformations between features and the pseudo labels and ignore the nonlinear relations. The second approach, "column subset selection", assumes only linear dependencies among the features and solves an optimization problem that is similar to principal component analysis (PCA) [16], [17]. Although methods in the third approach, such as [18], capture nonlinear relations, they ignore the interaction between the features.

Feature selection methods are usually classified in three main groups: wrappers, Filter and embedded [1]. In the wrapper method, the feature subsets are evaluated directly by an induction algorithm. In embedded methods, feature selection is performed during the training process of the given learning algorithm. Such approaches are usually computationally expensive and, hence, prohibitive in large data sets. An alternative solution is the *filter* approach in which an intermediate measure, independent of the induction learning algorithm, is used to evaluate the feature subsets. Filter methods are preferred as they are computationally more efficient and relatively robust against overfitting. The challenge in this area, that remains open, is to design a computationally efficient measure which is provably related to the generalization loss. Several measures has been introduced in the literature. Well-known criteria for feature selection can be grouped into similarity-based measures (e.g., Pearson correlation, Fisher Score), information-theoretic measures [11], [19]–[22], and Kernel-based measures [23]–[25]. Although correlation criteria are computationally more efficient, they usually are not able to detect *nonlinear* dependencies in features-label relations. Methods based on kernels can detect the nonlinear dependencies. However, the computational complexity of computing a kernel grows super linearly, if not quadratic, with the number of the samples [4]. Mutual Information (MI) criteria, on the other hand, can detect nonlinear dependencies with lower computational complexity [21]. In addition, mutual information can be used to bound the Bayes misclassification rate [26], [27]. However, estimating multi-variate mutual information is known to be a difficult task with high sample complexity.

Notations: We write $[m]$ for set $\{1, 2, \cdots, m\}$. For any subset $\mathcal{J} \subseteq [d]$ with ordered elements $\{j_1, j_2, \cdots, j_k\}$, the vectors $(X_{j_1}, X_{j_2}, \cdots, X_{j_k})$, and $(x_{j_1}, x_{j_2}, \cdots, x_{j_k})$ are denoted, respectively, by $\mathbf{X}^{\mathcal{J}}$ and $\mathbf{x}^{\mathcal{J}}$.

II. Fourier for Correlated Random Variables

In this section, we propose a novel Fourier expansion for functions of *correlated* binary features. For convenience in presenting our results, we restrict ourselves to binary features. One can extend our approach to discrete features via *group characters*.

We start with a brief overview of the well-known Fourier expansion on Boolean cube [5]. Let $\mathbf{X} = (X_1, X_2, ..., X_d)$ be a vector of mutually independent random variables taking values from a subset $\mathcal{X} \subset \mathbb{R}^d$. Let μ_j and σ_j be the mean and standard-deviation of $X_j, j \in [d]$. Suppose that these random variables are non-trivial, that is $\sigma_j > 0$ for all $j \in [d]$. The Fourier expansion is defined via a set of basis functions called *parities*. The *parity* for a subset $\mathcal{S} \subseteq [d]$ is defined as

$$\phi_{\mathcal{S}}(\mathbf{x}) \triangleq \prod_{i \in \mathcal{S}} \frac{x_i - \mu_i}{\sigma_i}, \qquad \text{for all } \mathbf{x} \in \mathbb{R}^d.$$

Since X_i's are mutually independent, the parities are orthonormal, that is $\mathbb{E}[\phi_{\mathcal{S}}(\mathbf{X})^2] = 1$ for any subset $\mathcal{S}$, and $\mathbb{E}[\phi_{\mathcal{S}}(\mathbf{X})\ \phi_{\mathcal{T}}(\mathbf{X})] = 0$ when $\mathcal{T} \neq \mathcal{S}$ (that is $\exists x \in \mathcal{T} \bigcup \mathcal{S}$ such that $x \notin \mathcal{T} \bigcap \mathcal{S}$). Under the assumption that $\mathcal{X} = \{-1, 1\}^d$, the parities form an orthonormal basis for the space of bounded function $f : \{-1, 1\}^d \mapsto \mathbb{R}$ [5]. That is, any bounded function $f : \{-1, 1\}^d \mapsto \mathbb{R}$ can be written as a linear combination of the form $f(\mathbf{x}) = \sum_{\mathcal{S} \subseteq [d]} \mathtt{f}_{\mathcal{S}}\ \phi_{\mathcal{S}}(\mathbf{x})$, for all $\mathbf{x} \in \{-1, 1\}^d$, where $\mathtt{f}_{\mathcal{S}} \in \mathbb{R}$ are called the *Fourier coefficients* of f with respect to $P_{\mathbf{X}}$, the distribution of $\mathbf{X}$. Further, the Fourier coefficients can be computed as $\mathtt{f}_{\mathcal{S}} = \mathbb{E}[f(\mathbf{X})\phi_{\mathcal{S}}(\mathbf{X})]$, for all subsets $\mathcal{S} \subseteq [d]$.

With this overview, we are ready to construct our Fourier expansion. Note that, in a general probability space with correlated features, the standard Fourier expansion is no longer well-defined. Because, the parities $\phi_{\mathcal{S}}$ are not necessarily orthogonal. That said, we construct our Fourier expansion by adopting a Gram-Schmidt-type procedure to make the parities

orthogonal. Then, we use this basis to develop our Fourier expansion for function of correlated random variables. The orthogonalization process is explained in the following.

A. Orthogonalization process:

Fix the following ordering for subsets of $[d]$:

$$\emptyset, \{1\}, \{2\}, \{1,2\}, \{3\}, \{1,3\}, \{2,3\}, \cdots, \{1,2,...,d\}. \quad (1)$$

For any pair of functions g_1, g_2 denote $\langle g_1, g_2 \rangle = \mathbb{E}[g_1(\mathbf{X})g_2(\mathbf{X})]$. We apply the Gram-Schmidt process on the parities $\phi_{\mathcal{S}_i}$ with the above ordering and $\langle g_1, g_2 \rangle$ as the inner product. With this method, the orthogonalized parity corresponding to the ith subset is obtained from the following

$$\tilde{\psi}_{\mathcal{S}_i} = \phi_{\mathcal{S}_i} - \sum_{j=1}^{i-1} \langle \psi_{\mathcal{S}_j}, \phi_{\mathcal{S}_i} \rangle \quad (2)$$

with $\psi_{\mathcal{S}_i} = \frac{\tilde{\psi}_{\mathcal{S}_i}}{\|\tilde{\psi}_{\mathcal{S}_i}\|_2}$ where $\|\tilde{\psi}_{\mathcal{S}_i}\|_2 = \sqrt{\langle \tilde{\psi}_{\mathcal{S}_i}, \tilde{\psi}_{\mathcal{S}_i} \rangle}$. Note that the first orthogonalized parity is given by $\psi_\emptyset(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathbb{R}^d$. By construction, the resulted nontrivial parities $\psi_{\mathcal{S}_i}$'s are orthonormal, that is $\langle \psi_{\mathcal{S}_i}, \psi_{\mathcal{S}_j} \rangle = 0$ for $i \neq j$ and $\langle \psi_{\mathcal{S}_i}, \psi_{\mathcal{S}_i} \rangle = 1$ if $\psi_{\mathcal{S}_i}$ is not trivial.

Depending on the statistics of the features, the number of non-trivial parities ranges from 1 to 2^d. On one extreme, if the features are mutually independent, then $\psi_{\mathcal{S}_i} = \phi_{\mathcal{S}_i}$. On the other extreme, if the features are trivial, then $\psi_{\mathcal{S}_i} = 0$ for $i > 1$, and hence there is only one non-trivial parity. Note also that different orderings for the subsets of $[d]$ result in different orthogonalized parities. We can show that ordering (1) is beneficial to remove "redundant" features. Hence, unless otherwise stated, we use the ordering in (1).

In the next proposition we establish our Fourier expansion for functions of correlated binary random variables.

Proposition 1 (Correlated Fourier Expansion). *Let $P_{\mathbf{X}}$ be any probability distribution on $\{-1,1\}^d$ and $f : \{-1,1\}^d \mapsto \mathbb{R}$ be a bounded function. Let $\psi_{\mathcal{S}}$'s be the orthogonalized parities as defined in (2). Then, for all $\mathbf{x} \in \{-1,1\}^d$ except a measure-zero subset, $f(\mathbf{x})$ is decomposed as $f(\mathbf{x}) = \sum_{\mathcal{S}\subseteq[d]} f_{\mathcal{S}}\psi_{\mathcal{S}}(\mathbf{x})$, where the summation is taken over all $\mathcal{S} \subseteq [d]$ for which $\psi_{\mathcal{S}}$ is not trivial. Further, the coefficients $f_{\mathcal{S}}$ are unique and obtained from $f_{\mathcal{S}} = \mathbb{E}[f(\mathbf{X})\psi_{\mathcal{S}}(\mathbf{X})]$.*

Example 1. *Set $d = 3$ and let X_1 and X_2 be independent random variables with Gaussian distribution $N(0,1)$. Suppose $X_3 = X_1X_2$ with probability one. There are eight standard parities, one for each subsets, as $(1, x_1, x_2, x_1x_2, x_3, x_1x_3, x_2x_3, x_1x_2x_3)$. By performing the orthogonalization process, as in (2), there are only four non-trivial orthogonalized parities as $\psi_\emptyset = 1, \psi_{\{1\}} = x_1, \psi_{\{2\}} = x_2$, and $\psi_{\{1,2\}} = x_1x_2$. The rest of the parities are zero, because $\|\tilde{\psi}_{\mathcal{S}}\|_2 = 0$ for any of the subsets $\{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}$. Now, suppose we change the relation of X_3 to $X_3 = \mathsf{sign}[X_1X_2]$. In this case, after the orthogonalization process, there are eight non-trivial parities. For instance, it is not difficult to check that $\tilde{\psi}_{\{3\}} = x_3 - \frac{2}{\pi}x_1x_2$. Hence, $\|\tilde{\psi}_{\{3\}}\|_2 > 0$, implying that $\psi_{\{3\}}$ is not redundant.*

This example shows that the orthogonalization removes nonlinear redundancies. However, it captures only a class of non-linearities for non-binary features. This is because the orthogonalization process is based on the binary Fourier expansion. We view our binary Fourier as a framework that captures a special class of nonlinearities — those characterized via orthogonalized parities. Our numerical experiments presented in Section V confirm that such an approximation is sufficient to outperform state-of-the-art methods for many data sets (see Table II). It is also noted that dimension reduction methods such as PCA do not necessarily capture the nonlinear redundancies. For instance, the features in the above example are pairwise uncorrelated and, hence, the covariance matrix is the identity matrix.

III. Information Sufficiency

We build upon our orthogonalization process in (2) and develop our UFFS algorithm (see Algorithm 1) to capture non-linear redundancies in the features. For this purpose, we first define a measure to identify the features as "sufficiently informative" and "redundant". Intuitively, the former group contains all the *information* accessible from the features. The later consists of the features that are a function of the "informative" features, and hence, can be removed from the data set.

Suppose that there are d features denoted by the random vector $\mathbf{X} = (X_1, X_2, ..., X_d)$ taking values from a subset $\mathcal{X} \subset \mathbb{R}^d$. We say $\mathcal{J} \subseteq [d]$ is a "sufficiently informative" feature subset, if $H(\mathbf{X}) = H(X^{\mathcal{J}})$, where H is the Shannon entropy. This definition is related to the notion of Markov Blanket [20], as $\mathcal{J}$ is a Markov blanket for any feature in $\mathcal{J}^c$. Also, $\mathcal{J}$ being sufficiently informative immediately leads to $\mathcal{J}^c$ being redundant. Because, the condition $H(\mathbf{X}) = H(X^{\mathcal{J}})$ implies that there exists a mapping T, such that $X^{\mathcal{J}^c} = T(X^{\mathcal{J}})$, with probability one [27]. Hence, all the features not included in $\mathcal{J}$ can be removed. With this elimination, the dimension is reduced from d to $|\mathcal{J}|$. As there are multiple such $\mathcal{J}$'s, the objective is to find the smallest one[1]. Tolerating small amounts of imperfections, we formalize the above notion in the following.

Definition 1 (Sufficiently Informative). *For discrete features and $0 \leq \epsilon \leq 1$, a feature subset $\mathcal{J}$ is said to be ϵ-sufficiently informative, if $H(\mathbf{X}|X^{\mathcal{J}}) \leq \epsilon$. The feature subset $\mathcal{J}$ is sufficiently informative, if $H(\mathbf{X}|X^{\mathcal{J}}) = 0$. Such $\mathcal{J}$ is called minimal, if it has the minimum cardinality among all sufficiently informative feature subsets.*

Next, we make a connection between the above definition and the orthogonalization process in (2). We employ this process to extract a sufficiently informative feature subset. Fix the standard ordering as in (1), and generate the orthogonalized parities $\psi_{\mathcal{S}_i}$. We start by deriving an upper-bound on $H(\mathbf{X}|X^{\mathcal{J}})$ in terms of the orthogonalized parities in (2). Note

[1]The set of all features is a trivial example of a sufficiently informative feature subset.

that $H(\mathbf{X}|X^{\mathcal{J}}) = H(X^{\mathcal{J}^c}|X^{\mathcal{J}})$, where $\mathcal{J}^c$ is the complement of $\mathcal{J}$. Thus, from the chain rule [27], this quantity equals to $\sum_{i\in\mathcal{J}^c} H(X_i|X^{\mathcal{J}}, X^{i-1})$. As X_i is a discrete random variable and $\phi_{\{i\}}(\mathbf{x}) = (x_i - \mu_i)/\sigma_j$, then

$$H(X_i|X^{\mathcal{J}}, X^{i-1}) = H\big(\phi_{\{i\}}(\mathbf{X})|X^{\mathcal{J}}, X^{i-1}\big),$$

Since $\phi_{\{i\}}$ is the standard parity as in Section II, then, from the orthogonalization process in (2), we can write $\phi_{\{i\}} = \tilde{\psi}_{\{i\}} + \sum_{\mathcal{S}\subseteq[i-1]} \alpha_{\mathcal{S}}\psi_{\mathcal{S}}$, where $\alpha_{\mathcal{S}} = \langle\phi_{\{i\}}, \psi_{\mathcal{S}}\rangle$. In this decomposition, the terms in the summation depend only on X^{i-1}. This is due to the particular ordering in (1). Therefore, we get the following upper-bound

$$H(\phi_{\{i\}}|X^{\mathcal{J}}, X^{i-1}) = H(\tilde{\psi}_{\{i\}}|X^{\mathcal{J}}, X^{i-1}) \leq H(\tilde{\psi}_{\{i\}}), \quad (3)$$

where the last inequality follows by removing the conditioning in the entropy. Lastly, adapting this bound for all $i \in \mathcal{J}^c$, we get the upper-bound: $H(\mathbf{X}|X^{\mathcal{J}}) \leq \sum_{i\in\mathcal{J}^c} H(\tilde{\psi}_{\{i\}})$. Using this bound, we can prove [7] the following theorem.

Theorem 1. *Let $\mathcal{J}_\epsilon \subseteq [d]$ be the set of all i's such that $\|\tilde{\psi}_{\{i\}}\|_2 > \epsilon$. Then, for sufficiently small $\epsilon > 0$, $H(\mathbf{X}|X^{\mathcal{J}_\epsilon}) = O(\epsilon)$. Further, if the features take values from $\{-1, 1\}^d$, then there exists an ordering of the features so that $\mathcal{J}_\epsilon$ with $\epsilon = 0$ is a sufficiently informative subset with minimum cardinality.*

Proof. From the discussion before the statement of the theorem and (3) with $\mathcal{J}$ replaced by $\mathcal{J}_\epsilon$, we have that

$$H(X^d|X^{\mathcal{J}_\epsilon}) \leq \sum_{i\in\mathcal{J}_\epsilon^c} H(\tilde{\psi}_{\{i\}}). \quad (4)$$

Let

$$a_i = \min\{|\tilde{\psi}_{\{i\}}(\mathbf{x})| : \mathbf{x} \in \mathcal{X}^d, \tilde{\psi}_{\{i\}}(\mathbf{x}) \neq 0\}, \quad \text{for all } i \in [d],$$

and define $a_{\min} \triangleq \min_i a_i$. From the assumption in the statement of the theorem, $\|\tilde{\psi}_{\{i\}}\|_2^2 \leq \epsilon^2$, for all $i \in \mathcal{J}_\epsilon^c$. Then, from Markov's inequality, we obtain that

$$P\{\tilde{\psi}_{\{i\}}(\mathbf{X}) \neq 0\} = P\{|\tilde{\psi}_{\{i\}}(\mathbf{X})|^2 \geq a_i^2\} \leq \frac{\epsilon^2}{a_i^2} \leq \frac{\epsilon^2}{a_{\min}^2}.$$

Therefore, from grouping axiom for entropy [27], we have that

$$H(\tilde{\psi}_{\{i\}}) \leq h_b(\frac{\epsilon^2}{a_{\min}^2}) + \frac{\epsilon^2}{a_{\min}^2}\log_2|\mathcal{X}|, \quad (5)$$

where $h_b(\cdot)$ is the binary entropy function. When $\epsilon \leq a_{min}$, by combining (4) and (5), the following inequality holds:

$$H(X^d|X^{\mathcal{J}_\epsilon}) \leq (d - |\mathcal{J}_\epsilon|)\Big(h_b(\frac{\epsilon^2}{a_{min}^2}) + \frac{\epsilon^2}{a_{min}^2}\log_2|\mathcal{X}|\Big). \quad (6)$$

Note that the binary entropy satisfies the inequality $h_b(p) \leq 2\sqrt{p(1-p)}$. Hence, for fixed d, a_{min} and $\mathcal{X}$, the right-hand side of the above inequality is $dO(\epsilon)$ as $\epsilon \to 0$. This completes the proof for the first statement of the theorem.

Next, we prove the second statement: "if the features take values from $\{-1, 1\}^d$, then $\mathcal{J}_\epsilon$ with $\epsilon = 0$ is a *sufficiently informative* subset with minimum cardinality." Note that, from Definition 1, the subset $\mathcal{J}_0$ is sufficiently informative. This is because for any $j \notin \mathcal{J}_0$ the parity $\tilde{\psi}_{\{j\}}(\mathbf{X}) = 0$ with probability one. Therefore, from (2), the standard parity $\phi_{\{j\}}$ is a function of $(X_1, X_2, ..., X_{j-1})$. Implying that X_j is a function of $(X_1, X_2, ..., X_{j-1})$. Hence, $H(X^d|X^{\mathcal{J}_0}) = 0$. It remains to prove that $\mathcal{J}_0$ is minimal when the features take values from $\{-1, 1\}^d$. We proceed by contradiction. Suppose $\mathcal{J}_0$ is not minimal. Then, there exists a sufficiently informative features subset $\mathcal{A}$ such that $|\mathcal{A}| < |\mathcal{J}_0|$. Consider the Hilbert space of bounded functions $f : \{-1, 1\}^d \mapsto \mathbb{R}$ with inner product defined as $\langle f, g\rangle = \mathbb{E}[f(\mathbf{X})g(\mathbf{X})]$. Since $\mathcal{A}$ is sufficiently informative, there exists a mapping Γ such that $X^d = \Gamma(X^{\mathcal{A}})$ with probability one. Therefore, any function $f(X^d)$ can be written as $f(X^d) = f(\Gamma(X^{\mathcal{A}}))$. This implies that the dimension of the Hilbert space is at most $2^{|\mathcal{A}|}$. However, there are $|\mathcal{J}_0|$ non-trivial parities $\psi_{\mathcal{S}_i}$ that, from Proposition 1, form an orthonormal basis for this Hilbert space. Hence, the dimension of this space cannot be less than $2^{|\mathcal{J}_0|}$. This is a contradiction and completes the proof.

IV. Unsupervised Learning Algorithm

As a result of Theorem 1, $\|\tilde{\psi}_{\{i\}}\|_2$ can be viewed as a measure of the *redundancy* of each feature and that the orthogonalization procedure can remove them. We use this measure for unsupervised feature selection, where n independent and identically distributed (i.i.d.) instances $\{\mathbf{x}(i), i \in [n]\}$ are available. The idea is to perform the orthogonalization process as in (2) and find the features j for which $\|\tilde{\psi}_{\{j\}}\|_2$ is smaller than a threshold ϵ. These features are declared as redundant. As for the algorithm, two issues need to be addressed: 1) the orthogonalization is an NP hard process, as there are 2^d feature subsets, and 2) estimation of $\|\tilde{\psi}_{\{j\}}\|_2$ from the training instances. In what follows, we address these issues.

a) **Fixed-depth search::** We propose to address the first issue using a *fixed-depth* search method. Given a parameter $t \leq d$, the orthogonalization is performed only on feature subsets of size at most t. For that we use the standard ordering as in (1), but restricted to subsets of size at most t.

b) **Empirical orthogonalization::** We propose a recursive formula to perform the orthogonalization and estimate $\|\tilde{\psi}_{\{j\}}\|_2$. Let $b_{j,i} = \langle\phi_{\mathcal{S}_j}, \phi_{\mathcal{S}_i}\rangle$, and define $a_{j,i} = \langle\psi_{\mathcal{S}_j}, \phi_{\mathcal{S}_i}\rangle$. Therefore, (2) can be written as that $\tilde{\psi}_{\mathcal{S}_i} = \phi_{\mathcal{S}_i} - \sum_{j<i} a_{j,i}\psi_{\mathcal{S}_j}$. Due to the orthonormality of $\psi_{\mathcal{S}_i}$'s, we obtain that $\|\tilde{\psi}_{\mathcal{S}_i}\|_2^2 = b_{i,i} - \sum_{j<i} a_{j,i}^2$. Further, the coefficients $a_{j,i}$ can be calculated recursively as

$$a_{j,i} = \frac{1}{\sqrt{b_{j,j} - \sum_{r<j} a_{r,j}^2}}\Big(b_{j,i} - \sum_{\ell<j} a_{\ell,j}a_{\ell,i}\Big). \quad (7)$$

With this formulas, we first compute an empirical estimate of $b_{j,i}$'s, denoted by $\hat{b}_{j,i}$. Then, we compute an estimation of $a_{j,i}$'s (denoted by $\hat{a}_{j,i}$) by calculating (7) with $b_{j,i}$ and $a_{j,i}$ replaced by $\hat{b}_{j,i}$ and $\hat{a}_{j,i}$, receptively. Lastly, we obtain an empirical estimate of $\|\tilde{\psi}_{\mathcal{S}_i}\|_2$ by computing $\sqrt{\hat{b}_{i,i} - \sum_{j<i}\hat{a}_{j,i}^2}$.

c) **Clustering the features::** The above two processes are implemented in Algorithm 1. For large dimensional data sets, we can group the features into multiple clusters of approximately equal size (say m features). Then, we perform Algorithm 1 on each cluster, and remove the redundant features within it. With this approach, the computational complexity of UFFS algorithm with depth parameter t and cluster size m is $O(n\frac{d}{m}m^{2t})$. The parameters m and t are chosen independently of (n, d). For instance, we choose $t = 3$ and $m = 40$. As a result, we obtain a complexity linear in the size of the data set. We present our experimental results next.

Algorithm 1 Orthogonalization Algorithm

Input: n training samples $\mathbf{x}_i \in \mathbb{R}^d$, depth parameter $t \leq d$, and redundancy threshold $\epsilon \in (0, 1)$

Output: Features' measures $\mathsf{norm}(j), j = 1, 2, ...d$

Compute the empirical mean $\hat{\mu}_j$ and standard deviation $\hat{\sigma}_j$ of each feature.

Generate all subsets $\mathcal{S}_i \subseteq [d]$ with size at most t and with the standard ordering as in (1). Compute the matrix $\hat{\mathbf{B}}$ with elements:

$$\hat{b}_{j,i} \leftarrow \frac{1}{n}\sum_{l=1}^{n}\Big[\prod_{u\in\mathcal{S}_j}\frac{x_{lu}-\hat{\mu}_u}{\hat{\sigma}_u}\Big]\Big[\prod_{v\in\mathcal{S}_i}\frac{x_{lv}-\hat{\mu}_v}{\hat{\sigma}_v}\Big].$$

Set $\hat{\mathbf{A}} \leftarrow \hat{\mathbf{B}}$

for row j of $\hat{\mathbf{A}}$ **do**

  update the jth row: $\hat{\mathbf{A}}_{j,*} \leftarrow \hat{\mathbf{A}}_{j,*} - \sum_{\ell<j}\hat{a}_{\ell,j}\hat{\mathbf{A}}_{\ell,*}$

  Compute $\mathsf{norm}(\mathcal{S}_j) \leftarrow \sqrt{[\hat{b}_{j,j} - \sum_{r<j}\hat{a}^2_{r,j}]^+}$

  if $\mathsf{norm}(\mathcal{S}_j) \leq \epsilon$ **then**

    Set the jth row of $\hat{\mathbf{A}}$ zero: $\hat{\mathbf{A}}_{j,*} \leftarrow \mathbf{0}$

  else

    Normalize the jth row: $\hat{\mathbf{A}}_{j,*} \leftarrow \frac{\hat{\mathbf{A}}_{j,*}}{\mathsf{norm}(\mathcal{S}_j)}$

Declare all $j \in [d]$ with $\mathsf{norm}(j) \geq \epsilon$ as non-redundant.

V. Numerical Experiments

We now compare the performance our UFFS algorithm (Algorithm 1) with a number of well-known methods for unsupervised feature selection. We tested the algorithms on several real-world data sets as given in Table I. These data sets are benchmarks and taken from [2] and the UCI repository [28]. In addition, we generated five synthetic data sets: S1, S2, S3 are for unsupervised feature selection.

For unsupervised feature selection, we generated three data sets, denoted by S1, S2, and S3. Each data set has 30 features: 10 informative denoted by $(X_1, X_2, ..., X_{10})$, 10 nonlinear redundant $(X_{11}, X_{12}, ..., X_{20})$, and 10 linearly redundant $(X_{21}, X_{12}, ..., X_{30})$. The informative features are generated according to three distributions, one for each data set. The distribution for S1 is $N(0, \mathbf{I}_{10})$, for S2 is uniform distribution over $[-1, 1]^{10}$, and for S3 is uniform distribution over $\{-1, 1\}^{10}$. Each nonlinear redundant feature is generated from $X_j = 3X_{i_1}X_{i_2}X_{i_3}$, where $j = 11, 12, ..., 20$, and i_1, i_2, i_3 are randomly and uniformly selected from $\{1, 2, ..., 10\}$. The linearly redundant features are generated from $X_j = \sum_{l=1}^{5} a_{j,l}X_{i_l}$, where i_l's are selected randomly from $\{1, 2, ..., 10\}$ and $a_{j,l} \sim \text{Unif}(0, 1)$. We use the above redundancy model for each data set. For the sake of performance comparison, we add a labeling to the above data sets. However, the labels are not revealed to the algorithms. We generate a fixed but randomly generated labeling function $f(\mathbf{X})$ on $\mathbb{R}^{10}$. This function is the sign of the following randomly generated polynomial in $\mathbb{R}^{10}$:

$$f(\mathbf{x}) = \mathsf{sign}\Big[\prod_{1\leq j\leq 3}\big(b_{0,j} + \sum_{1\leq i\leq 10} b_{i,j}x_i\big)\Big],$$

where $b_{i,j} \sim \text{Unif}(0, 1)$ and mutually independent.

We compare the performance of UFFS with Laplacian Score (LS) [18], MCFC [29], UDFS [15] and NDFS [14] on the real and the synthetic data sets. The labels are not revealed to the algorithms, but used for measuring the performances. Features are randomly ordered, so that the initial ordering would not affect the experiments' outcomes. Contrary to other algorithms, UFFS does not rank the feature; instead it outputs a set of indices as the non-redundant features. We run UFFS three times: first with $t = 1, m = d$, second, with $t = 2$, $m = 50$ but on the selected features from the first run, and third, with $t = 3, m = 30$ but on the selected features from the second run. For each experiment, let k denote the number of the selected features by UFFS at the third run. For comparing the performance to the ranking algorithms, we select only the k features with the highest rank. Once the features are selected by each unsupervised algorithm, we reveal the samples of the selects features with the labels to a classifier and compute its prediction accuracy. A support vector machine (SVM) classifier with radial basis function as kernel is employed for all the studies. We perform a 5-fold cross validation using this classifier and on the entire data set.

Table II shows the average of the resulted classification accuracies for each algorithm. The second row is the resulted accuracy without any feature selection. The third row is k which is the number of non-redundant features declared by the UFFS. Observe that, in synthetic data sets, k is very close to 10 which is the actual number of non-redundant features. The resulted accuracy by the UFFS is very close or greater than the accuracy without feature selection which verifies that the removed features were redundant.

TABLE I
Properties of the tested data sets.

Data set	S1	S2	S3	E1	E2	USPS	Isolet	COIL20	Covertype	Australian	Musk	ALL AML
Features	30	30	30	20	20	256	617	1024	46	14	166	7128
Samples	1000	1000	1000	1000	1000	9298	1560	1440	581	690	467	72

TABLE II
Comparison of unsupervised algorithms.

	S1	S2	S3	USPS	Isolet	COIL20	Covertype	Australian	Musk	ALL AML
No FS	77.9	75.0	87.0	97.3	92.8	98.8	75.6	84.9	92.2	94.3
UFFS k	11	12	11	93	309	331	34	12	35	39
UFFS	**80.3**	**76.8**	**86.2**	**97.0**	**91.7**	**98.8**	**76.9**	**85.1**	**85.7**	**97.1**
LS	55.1	61.2	71.0	95.6	88.6	98.9	72.8	85.4	84.5	97.2
MCFC	56.6	59.0	65.8	93.9	90.1	94.0	72.3	84.8	84.2	95.9
UDFS	64.0	60.6	64.3	80.8	90.2	98.0	72.0	84.9	80	86.2
NDFS	64.5	57.5	67.8	97.0	92.4	99.3	74.2	84.8	84.9	97.1

ACKNOWLEDGEMENT

This work was supported in part by NSF Center on Science of Information Grants CCF-0939370 and NSF Grants CCF-1524312, CCF-2006440, CCF-2007238, and Google Research Award.

REFERENCES

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[2] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, p. 94, 2018.

[3] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A review of unsupervised feature selection methods," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 907–948, 2020.

[4] N. Cesa-Bianchi, Y. Mansour, and O. Shamir, "On the complexity of learning with kernels," in *Conference on Learning Theory*, 2015, pp. 297–325.

[5] R. O'Donnell, *Analysis of boolean functions*. Cambridge University Press, 2014.

[6] R. d. Wolf, *A Brief Introduction to Fourier Analysis on the Boolean Cube*, ser. Graduate Surveys. Theory of Computing Library, 2008, no. 1. [Online]. Available: http://www.theoryofcomputing.org/library.html

[7] M. Heidari, n. G. S. J. Sreedharan, and W. Szpankowski, "Feature selection via a fourier framework," 2020.

[8] W. Hoeffding, "A class of statistics with asymptotically normal distribution," *The Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 293–325, 1948. [Online]. Available: http://www.jstor.org/stable/2235637

[9] I. M. Sobol, "Sensitivity estimates for nonlinear mathematical models," *Mathematical modelling and computational experiments*, vol. 1, no. 4, pp. 407–414, 1993.

[10] G. Chastaing, F. Gamboa, C. Prieur *et al.*, "Generalized hoeffding-sobol decomposition for dependent variables-application to sensitivity analysis," *Electronic Journal of Statistics*, vol. 6, pp. 2420–2448, 2012.

[11] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of machine learning research*, vol. 5, no. Oct, pp. 1205–1224, 2004.

[12] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[13] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[14] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, ser. AAAI'12. AAAI Press, 2012, p. 1026–1032.

[15] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "L2, 1-norm regularized discriminative feature selection for unsupervised," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[16] C. Feng, C. Qian, and K. Tang, "Unsupervised feature selection by pareto optimization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3534–3541.

[17] H. Arai, C. Maung, K. Xu, and H. Schweitzer, "Unsupervised feature selection by heuristic search with provable bounds on suboptimality," in *Proceedings of the Thirtieth AAAI conference on Artificial Intelligence*, 2016, pp. 666–672.

[18] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, 2006, pp. 507–514.

[19] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural computing and applications*, vol. 24, no. 1, pp. 175–186, 2014.

[20] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab, Tech. Rep., 1996.

[21] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[22] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[23] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2005, pp. 63–77.

[24] J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan, "Kernel feature selection via conditional covariance minimization," in *Advances in Neural Information Processing Systems*, 2017, pp. 6946–6955.

[25] X. Wei, B. Cao, and P. S. Yu, "Nonlinear joint unsupervised feature selection," in *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016, pp. 414–422.

[26] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 259–266, 1994.

[27] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley-Interscience, 2006.

[28] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[29] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 333–342.