Combinatorics of Distance Covariance: Inclusion-Minimal Maximizers of Quasi-Concave Set Functions for Diverse Variable Selection

Praneeth Vepakomma[a,b,*], Yulia Kempner[c]

[a] *Rutgers University*
[b] *Motorola Solutions*
[c] *Holon Institute of Technology*

Abstract

In this paper we show that the negative sample distance covariance function is a quasi-concave set function of samples of random variables that are not statistically independent. We use these properties to propose greedy algorithms to combinatorially optimize some diversity (low statistical dependence) promoting functions of distance covariance. Our greedy algorithm obtains all the inclusion-minimal maximizers of this diversity promoting objective. Inclusion-minimal maximizers are multiple solution sets of globally optimal maximizers that are not a proper subset of any other maximizing set in the solution set. We present results upon applying this approach to obtain diverse features (covariates/variables/predictors) in a feature selection setting for regression (or classification) problems. We also combine our diverse feature selection algorithm with a distance covariance based relevant feature selection algorithm of [7] to produce subsets of covariates that are both relevant yet ordered in non-increasing levels of diversity of these subsets.

Keywords: Distance covariance, quasi-concave set function, minimal-maximizers, regression, diverse feature selection, greedy algorithm, combinatorics.

1. Introduction

1.1. The classical problem of variable selection:

The problem of "variable selection" also known as "feature selection" or "covariate selection" is a prominent problem in statistics and machine learning. The goal in here is to be able to choose an optimal subset of covariates in a regression or classification setting that would perform optimally with respect to the out-of-sample prediction or classification accuracy when the chosen subset is used to predict (or classify) one or more real-valued response

*corresponding author
Email addresses: praneeth@scarletmail.rutgers.edu (Praneeth Vepakomma), yuliak@hit.ac.il (Yulia Kempner)

variables (in regression) or one or more categorical variables (in classification). There have been an umpteen number of techniques developed for this problem under a broadly varying spectrum of assumptions.

1.2. The more recent problem of diverse variable selection:

Traditional feature selection algorithms have the primary goal of finding the best feature subset that is relevant to a regression or classification task. More recently, there has been a strong focus on not just the above mentioned goal of relevant feature selection but also on selecting a "small" subset of "diverse" features. Diversity is useful for several reasons such as interpretability, robustness to noise and in some cases to cater to reduction of real-life costs of costly feature acquisition for guiding feature engineering to decide on what other features could be acquired etc.

1.2.1. Some existing work on diversification:

The authors in [2] provide a solution in the specific case of linear regression through a formulation where a diversity promoting sub-modular regularizer is added to the standard linear regression problem. In this setting the solution is obtained by greedy algorithms that optimize a submodular function based objective. Although this is an interesting approach, we'd like to point that this approach restricts the regression model to be linear unlike it being generalized to any regression (non-linear and linear) models. Another important issue with this approach is that their solution is not globally optimal but is instead an approximation with a well-known (in submodular optimization) $1 - \frac{1}{e}$ styled guarantee of $\frac{(1-e^{-b.\gamma(U,k)}).OPT}{c}$, where OPT is the optimal solution, $\gamma(U, k)$ is a function of the solution subset of features U obtained through their algorithm and it's cardinality k (i.e., the number of features in U). b, c are algorithm dependent constants depending on the specific choice of *algorithm* out of multiple algorithms that they propose. Note that $(1 - \frac{1}{e})$ is approximately equal to 63%.

1.2.2. Existing work on diversification with mutual-information:

Another popular approach is [1] which is based on measuring diversity and relevancy through functions of mutual-information. Their solution approximately optimizes their proposed objective as obtaining a global solution would require $O(n^{|S|})$ search operations where n is the number of samples and $|S|$ is the cardinality of the number of features required to be selected by the algorithm. This can be a prohibitively large number in the case of many practical datasets and required $|S|$.

1.3. Advantages of our proposed algorithms:

The technique proposed in below sections of our paper has two major advantages:

1. Our solution to our proposed diversity encouraging objective function is globally optimal with no approximation error unlike the $1-\frac{1}{e}$ styled approximate solution provided by [2] or the unquantified approximation error provided by [1]. We also propose an approach that is completely devoid of any parameters and provide a global solution to our proposed formulation. That said, we do completely recognize that the objective

function proposed in our technique varies from the objective functions proposed in existing techniques.

2. Another advantage of our approach is that it is independent of the choice of regression (linear/non-linear) or classification (linear/non-linear) model to be used unlike the work by [2] which focusses only on linear regression.

3. Our approach can directly be used for diversified feature selection in both cases of univariate or multivariate (vector-valued) responses (in regression) or multi-label (in classification) without modifying our proposed objective function or algorithm while the approach in [2] does not seem to extend trivially beyond the univariate response case in linear regression without modifying their regularized objective function or algorithmic routines.

Prior to getting into the crux of our proposed theoretical results and algorithmic implications, we'd like to note that in theory our approach can be explicitly parametrized by a trade-off parameter to control the trade-off between relevancy and diversity of features selected. Such tuning of trade-offs is not the main focus of this paper. The previously proposed approach using spectral regularization [2] does parametrize this through regularization parameters that weigh the submodular regularizer appropriately.

2. Problem Formulation:

In this paper we cover the following three problems:

Problem I: **Diverse Feature Selection**
The goal here is to find a subset of features that have the least statistical dependence amongst each other. This implies that the selected features would be diverse.

Problem II: **All-Relevant Feature Selection**
The goal here is to find a subset of features that are most statistically dependent on a response variable.

Problem III: **Diverse and Relevant Feature Selection**
The goal here is to find a subset of features that are more statistically dependent on a response variable while also being less statistically dependent amongst each other.

We present a greedy-algorithm with exactly optimal solutions in this paper for our formulated objective to solve Problem I. We point to an existing approach for Problem II and propose simple methodologies for Problem III where the methodologies are based on solutions of Problem I and II. Before we get to the main result of our paper, our suggested two simple methodological approaches for Problem III are:

(a.) Controlled approach: In this approach, we first choose a subset of features that "individually" have a statistical dependency $i.e \geq \alpha \in \mathcal{R}^+$ with response variable and call this subset the controlled set. We then run our algorithm proposed for Problem I for choosing a diverse set of features from this controlled set.

(b.) Two-stage approach: In this approach, the Problem III could be approached by solving Problem II followed by Problem I or vice-versa.

2.1. ***Main Result of the paper:***

So to clearly reiterate, our main and most novel contribution of this paper is our proposed algorithm for Problem I.

3. Preliminaries:

In this section we introduce some preliminaries about distance correlation and distance covariance which we extensively use in our paper to build up towards our proposed theoretical results.

3.1. ***Distance Covariance and Distance Correlation:***

Distance Correlation [3] is a measure of nonlinear statistical dependencies between random vectors of arbitrary dimensions. We describe below distance covariance $\nu^2(\mathbf{x}, \mathbf{y})$ between random variables $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^m$ with finite first moments is a non-negative number as

$$\nu^2(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^{d+m}} |f_{\mathbf{x},\mathbf{y}}(t, s) - f_{\mathbf{x}}(t) f_{\mathbf{y}}(s)|^2 w(t, s) dt ds \tag{1}$$

where $w(t, s)$ is a weight function as defined in [3], $f_{\mathbf{x}}, f_{\mathbf{y}}$ are characteristic functions of $\mathbf{x}, \mathbf{y}$ and $f_{\mathbf{x},\mathbf{y}}$ is the joint characteristic function.

The distance covariance is zero if and only if random variables $\mathbf{x}$ and $\mathbf{y}$ are independent. Using the above definition of distance covariance, we have the following expression for Distance Correlation from [3]:

The squared Distance Correlation between random variables $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^m$ with finite first moments is a nonnegative number is defined as

$$\rho^2(\mathbf{x}, \mathbf{y}) = \begin{cases} \frac{\nu^2(\mathbf{x},\mathbf{y})}{\sqrt{\nu^2(\mathbf{x},\mathbf{x})\nu^2(\mathbf{y},\mathbf{y})}}, & \nu^2(\mathbf{x}, \mathbf{x})\nu^2(\mathbf{y}, \mathbf{y}) > 0. \\ 0, & \nu^2(\mathbf{x}, \mathbf{x})\nu^2(\mathbf{y}, \mathbf{y}) = 0. \end{cases} \tag{2}$$

The Distance Correlation defined above has the following interesting properties;

1. $\rho^2(\mathbf{x}, \mathbf{y})$ is applicable for arbitrary dimensions d and m of $\mathbf{x}$ and $\mathbf{y}$ respectively.
2. $\rho^2(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent.
3. $\rho^2(\mathbf{x}, \mathbf{y})$ satisfies the relation $0 \leq \rho^2(\mathbf{x}, \mathbf{y}) \leq 1$.

3.2. ***Sample Distance Covariance and Sample Distance Correlation:***

We provide the definition of sample version of distance covariance [3] given samples $\{(\mathbf{x}_k, \mathbf{y}_k)|k = 1, 2, \ldots, n\}$ sampled i.i.d. from joint distribution of random vectors $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^m$. To do so, we define two squared Euclidean distance matrices $\mathbf{E_X}$ and $\mathbf{E_Y}$, where each entry $[\mathbf{E_X}]_{k,l} = \|\mathbf{x}_k - \mathbf{x}_l\|^2$ and $[\mathbf{E_Y}]_{k,l} = \|\mathbf{y}_k - \mathbf{y}_l\|^2$ with $k, l \in \{1, 2, \ldots, n\}$. These squared distance matrices are then double-centered by making their row and column sums zero and are denoted as $\widehat{\mathbf{E}}_\mathbf{X}, \widehat{\mathbf{Q}}_\mathbf{X}$, respectively. So given a double-centering matrix $\mathbf{J} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, we have $\widehat{\mathbf{E}}_\mathbf{X} = \mathbf{J}\mathbf{E_X}\mathbf{J}$ and $\widehat{\mathbf{E}}_\mathbf{Y} = \mathbf{J}\mathbf{E_Y}\mathbf{J}$. The sample distance covariance and sample distance correlation can now be defined as follows:

Definition 3.1. Sample Distance Covariance [3]: Given i.i.d samples $\mathcal{X}\times\mathcal{Y} = \{(\mathbf{x}_k, \mathbf{y}_k)|k = 1, 2, 3, \ldots, n\}$ and corresponding double centered Euclidean distance matrices $\widehat{\mathbf{E}}_\mathbf{X}$ and $\widehat{\mathbf{E}}_\mathbf{Y}$, then the squared sample distance correlation is defined as,

$$\hat{\nu}^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} [\widehat{\mathbf{E}}_\mathbf{X}]_{k,l}[\widehat{\mathbf{E}}_\mathbf{Y}]_{k,l},$$

Using this, sample distance correlation is given by

$$\hat{\rho}^2(\mathbf{X}, \mathbf{Y}) = \begin{cases} \frac{\hat{\nu}^2(\mathbf{X},\mathbf{Y})}{\sqrt{\hat{\nu}^2(\mathbf{X},\mathbf{X})\hat{\nu}^2(\mathbf{Y},\mathbf{Y})}}, & \hat{\nu}^2(\mathbf{X}, \mathbf{X})\hat{\nu}^2(\mathbf{Y}, \mathbf{Y}) > 0. \\ 0, & \hat{\nu}^2(\mathbf{X}, \mathbf{X})\hat{\nu}^2(\mathbf{Y}, \mathbf{Y}) = 0. \end{cases}$$

4. Kosorok's Distance Covariance Independence Inequality:

If $\mathbf{X}, \mathbf{Z} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$ and if and only if $\mathbf{Z} \perp\!\!\!\perp (\mathbf{X}, \mathbf{Y})$ then

$$\nu^2(\mathbf{X} + \mathbf{Z}, \mathbf{Y}) \leq \nu^2(\mathbf{X}, \mathbf{Y}) \tag{3}$$

Note that $\perp\!\!\!\perp$ indicates 'statistically independent' in statistical literature. This implies that for each $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ that are not pairwise statistically independent (i.e distance covariance between components of any subset of cardinality 2 of $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ is positive) then

$$\nu^2(\mathbf{X} + \mathbf{Z}, \mathbf{Y}) > \nu^2(\mathbf{X}, \mathbf{Y}) \tag{4}$$

5. Proof of Kosorok's Distance Covariance Inequality

The Kosorok's Distance Covariance Independence Inequality was proved in [7, 5] and is based on the property of characteristic functions (denoted below by f) that

$$|f_{\mathbf{X}+\mathbf{Z},\mathbf{Y}}(t, s) - f_{\mathbf{X}+\mathbf{Z}}(t)f_\mathbf{Y}(s)|^2 \leq |f_\mathbf{Z}(t)|^2|f_{\mathbf{X},\mathbf{Y}}(t, s) - f_\mathbf{X}(t)f_\mathbf{Y}(s)|^2 \tag{5}$$

and $|f_\mathbf{Z}(t)|^2 \leq 1$. The equation (5) above can be obtained by these facts

$$\begin{aligned} |\, f_{\mathbf{X}+\mathbf{Z},\mathbf{Y}}(t, s) - f_{\mathbf{X}+\mathbf{Z}}(t)f_\mathbf{Y}(s)|^2 &= |\,\mathbb{E}\, e^{it^T(\mathbf{X}+\mathbf{Z})+is^T\mathbf{Y}} - \mathbb{E}\, e^{it^T(\mathbf{X}+\mathbf{Z})}\,\mathbb{E}\, e^{is^T\mathbf{Y}}|^2 \\ &= |\,\mathbb{E}\, e^{it^T\mathbf{X}+is^T\mathbf{Y}}\,\mathbb{E}\, e^{it^T\mathbf{Z}} - \mathbb{E}\, e^{it^T\mathbf{X}}\,\mathbb{E}\, e^{it^T\mathbf{Z}}\,\mathbb{E}\, e^{it^T\mathbf{Y}}|^2 \\ &= |f_{\mathbf{X},\mathbf{Y}}(t, s)f_\mathbf{Z}(t) - f_\mathbf{X}(t)f_\mathbf{Z}(t)f_\mathbf{Y}(s)|^2 \\ &= |f_\mathbf{Z}(t)|^2|f_{\mathbf{X},\mathbf{Y}}(t, s) - f_\mathbf{X}(t)f_\mathbf{Y}(s)|^2 \end{aligned} \tag{6}$$

which with implication from $|f_{\mathbf{Z}}(t)|^2 \leq 1$ gives

$$\nu^2(\mathbf{X}+\mathbf{Z},\mathbf{Y}) \leq \nu^2(\mathbf{X},\mathbf{Y}) \tag{7}$$

We know that if $\mathbb{E}\,|\mathbf{X}|_d < \infty$, $\mathbb{E}\,|\mathbf{X}+\mathbf{Z}|_m < \infty$ and $\mathbb{E}\,|\mathbf{Y}|_d < \infty$, then from [3]

$$\lim_{n\to\infty} \nu_n^2(\mathbf{X}+\mathbf{Z},\mathbf{Y}) = \nu^2(\mathbf{X}+\mathbf{Z},\mathbf{Y})$$

and

$$\lim_{n\to\infty} \nu_n^2(\mathbf{X},\mathbf{Y}) = \nu^2(\mathbf{X},\mathbf{Y})$$

Thus, for the sample distance covariance, if n is large enough, we should have

$$V_n^2(\mathbf{X}+\mathbf{Z},\mathbf{Y}) \leq V^2(\mathbf{X},\mathbf{Y})$$

only under the assumption of independence between $(\mathbf{X},\mathbf{Y})$ and $\mathbf{Z}$. Note that ν_n indicates sample distance covariance and ν indicates population distance covariance.

Note: In the case where considering $(\mathbf{X}\cup\mathbf{Z})$ is of interest, we could use the above theorem by incorporating degenerated random vectors as follows: Suppose $\mathbf{X} \in \mathbb{R}^{p1}$ and $\mathbf{Z} \in \mathbb{R}^{p2}$, then we augment $\mathbf{X}$ and $\mathbf{Z}$ to be $\tilde{\mathbf{X}} = (\mathbf{X}, \mathbf{0}_{p2})$ and $\tilde{\mathbf{Z}} = (\mathbf{0}_{p1}, \mathbf{Z})$ respectively. $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Z}}$ are therefore of the same dimension and $\tilde{\mathbf{X}} + \tilde{\mathbf{Z}} = (\mathbf{X}, \mathbf{Z})$. Therefore the $\mathbf{X}\cup\mathbf{Z}$ operation in the context of computing $\hat{\nu}(\mathbf{X}\cup\mathbf{Z},\mathbf{Y})$ with matrices $\mathbf{X},\mathbf{Z},\mathbf{Y}$ is equivalent to appending the columns of $\mathbf{X}$ with the columns of $\mathbf{Z}$ followed by computing the sample-distance covariance between the resulting matrix and $\mathbf{Y}$.

6. Quasi-Concave Set Functions

6.1. Notation and definitions:

We now describe some notation and introduce some definitions that we use through out the paper in the sections below. We use bold faced $\mathbf{X}$ to denote the complete ground set of features/covariates and indexed X_i to denote the i'th covariate. That is we use i indexed subsets like S_i to indicate a singleton (unit cardinality) element of $\mathbf{S}$ labeled by i. We denote the response variable in a regression setting with $\mathbf{Y}$. We denote the set $2^{\mathbf{X}} \setminus \{\phi, \mathbf{X}\}$ by $\mathcal{P}^{-X}$ and use $\setminus$ to denote set difference, i.e $\mathbf{X} \setminus \mathbf{Z} = \{x : x \in \mathbf{X} \text{ and } x \notin \mathbf{Z}\}$.

Given a set system $(\mathbf{X},\mathcal{F})$ which is a collection $\mathcal{F}$ of subsets of a ground set $\mathbf{X}$ where $\mathcal{F} \subseteq 2^{\mathbf{X}}$, we define a quasi-concave set function as given below.

Definition 6.1 (**Quasi-Concave Set Function [4],[9]:**). A function $F : \mathcal{F} \mapsto \mathbb{R}$ defined on a set system $(\mathbf{X},\mathcal{F})$ is quasi-concave if for each $\mathbf{S},\mathbf{T} \in \mathcal{F}$,

$$F(\mathbf{S}\cap\mathbf{T}) \geq \min\{F(\mathbf{S}), F(\mathbf{T})\} \tag{8}$$

Definition 6.2 (**Monotone Linkage Function [9]:**). A function $\pi(X_i,\mathbf{Z})$ defined on $\mathbf{Z} \in \mathcal{P}^{-X}, X_i \in \mathbf{X}\setminus\mathbf{Z}$ is called a monotone linkage function if

$$\pi(X_i,\mathbf{S}) \geq \pi(X_i,\mathbf{T}), \mathbf{S}\subseteq\mathbf{T} \in \mathcal{F}, \forall X_i \in \mathbf{X}\setminus T \tag{9}$$

We'd like to note for the clarity of the reader that X_i is an element while $\mathbf{S},\mathbf{T}$ are sets. Therefore, to make this distinction clear we denote sets in bold-faced font and elements otherwise.

7. Some Combinatorial Properties of Negative Distance Covariance

We now prove some quasi-concave as well as monotone linkage set function properties of some functions of negative distance covariance.

Theorem 7.1 (Quasi-Concave Distance Covariance Set Function Theorem)**.** *If we have* $\mathbf{S} \cap \mathbf{T} \neq \varnothing$ *and* $\forall \mathbf{S}, \mathbf{T}, \mathbf{Y}$ *if* $\nu^2(\mathbf{S}, \mathbf{T}) > 0 \wedge \nu^2(\mathbf{S}, \mathbf{Y}) > 0 \wedge \nu^2(\mathbf{T}, \mathbf{Y}) > 0$ *then we have*

$$-\nu^2(\mathbf{S} \cap \mathbf{T}, \mathbf{Y}) \geq min(-\nu^2(\mathbf{S}, \mathbf{Y}), -\nu^2(\mathbf{T}, \mathbf{Y})) \tag{10}$$

Proof. If $\mathbf{S} \cap \mathbf{T} = \mathbf{S}$ then since $\mathbf{S} \subseteq \mathbf{T}$
the Kosorok's distance covariance inequality implies

$$-\nu^2(\mathbf{S}, \mathbf{Y}) \geq -\nu^2(\mathbf{T}, \mathbf{Y}) \tag{11}$$

Therefore we have

$$-\nu^2(\mathbf{S} \cap \mathbf{T}, \mathbf{Y}) \geq min(-\nu^2(\mathbf{S}, \mathbf{Y}), -\nu^2(\mathbf{T}, \mathbf{Y}))$$

Similarly, if $\mathbf{S} \cap \mathbf{T} = \mathbf{T}$, then since $\mathbf{T} \subseteq \mathbf{S}$

$$-\nu^2(\mathbf{T}, \mathbf{Y}) \geq -\nu^2(\mathbf{S}, \mathbf{Y}) \tag{12}$$

and therefore

$$-\nu^2(\mathbf{S} \cap \mathbf{T}, \mathbf{Y}) \geq min(-\nu^2(\mathbf{S}, \mathbf{Y}), -\nu^2(\mathbf{T}, \mathbf{Y})) \tag{13}$$

In the cases of $\mathbf{S} \cap \mathbf{T} \subset \mathbf{S}$ and $\mathbf{S} \cap \mathbf{T} \subset \mathbf{T}$ the Kosorok's distance covariance inequality implies

$$-\nu^2(\mathbf{S} \cap \mathbf{T}, \mathbf{Y}) > -\nu^2(\mathbf{S}, \mathbf{Y}) \tag{14}$$

and

$$-\nu^2(\mathbf{S} \cap \mathbf{T}, \mathbf{Y}) > -\nu^2(\mathbf{T}, \mathbf{Y}) \tag{15}$$

So

$$-\nu^2(\mathbf{S} \cap \mathbf{T}, \mathbf{Y}) \geq min(-\nu^2(\mathbf{S}, \mathbf{Y}), -\nu^2(\mathbf{T}, \mathbf{Y})) \tag{16}$$

□

7.1. ***A monotone linkage function of distance covariance:***

Lemma 7.2. *The function* $\pi(X_i, \mathbf{S})$ *of distance covariance defined on* $X_i \notin \mathbf{S}$ *as*

$$\underset{X_i \notin \mathbf{S}}{\pi(X_i, \mathbf{S})} = \sum_{\mathbf{S}_j \in \mathbf{S}} -\nu^2(X_i, \mathbf{S}_j) \tag{17}$$

is a monotone linkage function

Proof: For $\mathbf{S} \subseteq \mathbf{T}$ we have

$$\underset{X_i \notin \mathbf{T}}{\pi(X_i, \mathbf{T})} = \sum_{\mathbf{S}_j \in \mathbf{S}} -\nu_i^2(X_i, \mathbf{S}_j) - \sum_{\mathbf{T}_j \in \mathbf{T} \setminus \mathbf{S}} \nu_i^2(X_i, \mathbf{T}_j) \leq \underset{X_i \notin \mathbf{T}}{\pi(X_i, \mathbf{S})} = \sum_{\mathbf{S}_j \in \mathbf{S}} -\nu_i^2(X_i, \mathbf{S}_j) \tag{18}$$

We would also like to note that as $\nu(\cdot)$ is a non-negative function the above inequality does hold true.

Theorem 7.3. *[4]*

The function $M_\pi(\mathbf{T}) = \min_{X_i \in \mathbf{X} \setminus \mathbf{T}} \pi(X_i, \mathbf{T})$ *is a quasi-concave set function.*

Proof: The proof is in the proof of Assertion 1 in [4]

8. Diverse Feature Selection:

We aim to find all the subsets that maximize the function $M_\pi(\mathbf{T})$ which result in the solutions which for diverse features.

$$\arg \max_{\mathbf{T} \subset \mathbf{X}} \; M_\pi(\mathbf{T}) \tag{19}$$

The above equation (19) can be written as

$$\arg \max_{\mathbf{T} \subset \mathbf{X}} \; \min_{X_i \in \mathbf{X} \setminus \mathbf{T}} \pi(X_i, \mathbf{T}) \tag{20}$$

This problem does not necessarily have a single, unique solution and hence we aim to find all the subsets that are maximizers of (20). These are essentially subsets that are each maximally separated from their corresponding nearest neighbor where the notion of nearness to their neighbor is given by (17).

Definition 8.1 (π**-series:**)**.** We refer to a series $s_\pi = (X_{i_1}, \ldots, X_{i_N})$ as a π-series if

$$\pi(X_{i_{k+1}}, \overline{\mathbf{S}}_\mathbf{k}) = \min_{\mathbf{X_i} \in \mathbf{X} \setminus \overline{\mathbf{S}}_\mathbf{k}} \pi(\mathbf{X_i}, \overline{\mathbf{S}}_\mathbf{k}) \tag{21}$$

for any starting set $\overline{\mathbf{S}}_\mathbf{k} = \{\mathbf{X_{i_1}}, \ldots, \mathbf{X_{i_k}}\}, \mathbf{k} = \mathbf{1}, \ldots, \mathbf{N-1}$.

Therefore it is a way of greedily populating a series that can start with any first element $\mathbf{X_{i_1}}$ being the current series, but the subsequent element to be added to the series, must be the element that minimizes the element to current series function of $\pi(\mathbf{X_{i_{k+1}}}, \overline{\mathbf{S}}_\mathbf{k})$ where $\mathbf{X_{i_{k+1}}}$ is the next element added and $\overline{\mathbf{S}}_\mathbf{k}$ is the current series.

Definition 8.2 (π**-cluster**)**.** A subset $\mathbf{S} \in \mathcal{P}^{-\mathbf{X}}$ will be referred to as a π-cluster if there exists a π-series, $s_\pi = (X_{i_1}, \ldots, X_{i_N})$, such that $\mathbf{S}$ is a maximizer of $M_\pi(\overline{\mathbf{S}}_\mathbf{k})$ over all starting sets $\overline{\mathbf{S}}_\mathbf{k}$ of s_π.

Theorem 8.1. *[4] If for a π-series $s_\pi = (X_{i_1}, X_{i_2}, \ldots, i_N)$, a subset $\mathbf{S} \subset \mathbf{X}$ contains X_{i_1}, and if $X_{i_{k+1}}$ is the first element in s_π not contained in $\mathbf{S}$ (for some $k \in \{1, \ldots, N-1\}$, then*

$$M_\pi(\overline{\mathbf{S}}_\mathbf{k}) \geq M_\pi(\mathbf{S}) \tag{22}$$

where $\overline{\mathbf{S}}_\mathbf{k} = (X_{i_1}, \ldots, X_{i_k})$. In particular, if $\mathbf{S}$ is an inclusion-minimal maximizer of M_π (with regard to $\mathcal{P}^{-\mathbf{X}}$), then $\mathbf{S} = \overline{\mathbf{S}}_\mathbf{k}$, that is, $\mathbf{S}$ is a π-cluster.

Proof. $M_\pi(\overline{\mathbf{S}}_\mathbf{k}) = \pi(\mathrm{X}_{i_{k+1}}, \overline{\mathbf{S}}_\mathbf{k})$ by definition. Since $\overline{\mathbf{S}}_\mathbf{k} \subseteq \mathbf{S}$ we have $\pi(X_{i_{k+1}}, \overline{\mathbf{S}}_\mathbf{k}) \geq \pi(\mathrm{X}_{\mathbf{i_{k+1}}}, \mathbf{S})$ by monotonicity. To end the proof, note that $\pi(X_{i_{k+1}}, \mathbf{S}) \geq \mathbf{M}_\pi(\mathbf{S})$ because $M_\pi(\mathbf{S}) = \min_{X_i \in \mathbf{X} \setminus \mathbf{Z}} \pi(X_i, \mathbf{S})$ and $X_{i_{k+1}} \notin \mathbf{S}$ □

Proposition 8.2. *[4] If $\mathbf{S_1}, \mathbf{S_2} \subset \mathbf{X}$ are overlapping maximizers of a quasi-concave set function $M_\pi(\mathbf{S})$ over $\mathcal{P}^{-\mathbf{X}}$, then $\mathbf{S_1} \cap \mathbf{S_2}$ is also a maximizer of $M_\pi(\mathbf{S})$.*

Proof. It directly follows from (8). □

This implies that the minimal maximizers of a quasi-convex set function are not overlapping. Moreover, any nonminimal maximizer can be uniquely partitioned into a set of the minimal ones.

Theorem 8.3. *Each maximizer of a quasi-concave set function on $\mathcal{P}^{-\mathbf{X}}$ is a union of its inclusion-minimal maximizers.*

Proof. Indeed, if $\mathbf{S}^*$ is a maximizer of $M_\pi(\mathbf{S})$ over $\mathcal{P}^{-\mathbf{X}}$, then, according to Theorem 8.1, for any $X_i \in \mathbf{S}^*$, there exists a minimal maximizer included in $\mathbf{S}^*$ and containing X_i. □

8.1. ***Our greedy algorithm for diverse variable selection with distance covariance for solving Problem I:***

Algorithm 1 DiverseMinimalMaximDCoV: Diverse Combinatorial Distance Covariance

1: **function** =DiverseMinimalMaximDCoV($\mathbf{X}$)
2: | **for all** $X_i \in \mathbf{X}$ **do**
3: | Greedily form π-series $s_\pi(x) = (X_i, X_{i_2} \dots X_{i_N})$ starting from X_i as its first element.
4: | | **for** each π-series $s_\pi(x)$ in step 3 **do**
5: | Find a corresponding smallest starting subset $\mathbf{T_x}$ with

$$M_\pi(\mathbf{T_x}) = \max_{\mathbf{1 \le k \le N-1}} \pi(\mathbf{X_{i_{k+1}}}, \{\mathbf{X_{i_1}}, \dots, \mathbf{X_{i_k}}\})$$

6: | | **end for**
7: | **end for**
8: | Among the non-coinciding minimal π-clusters T_x's choose those that maximize

$$M_\pi(\mathbf{T_x}) = \min_{\mathbf{X_i} \in \mathbf{X} \setminus \mathbf{T_x}} \pi(\mathbf{X_i}, \mathbf{T_x})$$

all of which are the required minimal maximizers, and we return them as minimalMax
9: **return** (minimalMax)
10: **end function**

The above algorithm finds all minimal maximizers in $\mathcal{O}(N^3 g)$ time where g is the average time required to compute the value of $\pi(X_i, \mathbf{S})$ for any $X_i, \mathbf{S}$. The fastest version of computing distance covariance to date is $\mathcal{O}(NlogN)$ and proposed in [6].

Theorem 8.4. *The algorithm above finds all the minimal maximizers over* $\mathcal{P}^{-\mathbf{X}}$.

Proof. From Theorem 8.1 it follows that each element of minimalMax is a maximizer of $M_\pi(\mathbf{S})$ over $\mathcal{P}^{-\mathbf{X}}$. Assume that there is a minimal maximizer $\mathbf{S}$ that does not belong to minimalMax, and let $X_i \in \mathbf{S}$. Then, according to Theorem 8.1, there exist π-series starting from X_i and minimal π-cluster $T_x \subseteq \mathbf{S}$ containing X_i with $M_\pi(\mathbf{T_x}) \geq \mathbf{M}_\pi(\mathbf{S})$. Since $\mathbf{S}$ does not belong to minimalMax, and, according to steps 5 and 8 of the algorithm, T_x or some subset of T_x belongs to minimalMax, there are a minimal maximizer strictly included in $\mathbf{S}$ which contradicts the minimality of $\mathbf{S}$. □

Putting all these results together we present our algorithm in Algorithm 1 above.

9. All-Relevant Feature Selection

In addition to our algorithm proposed above, we would like to point to a recent algorithm proposed in [7] for the purpose of solving Problem II using distance covariance. We present this algorithm in Algorithm 2 below.

Algorithm 2 Kong-Wang-Wahba's All-Relevant Feature Selection algorithm for Problem II:

function KONG-WANG-WAHBA'S ALGORITHM($\mathbf{X}$)
  Calculate marginal sample distance correlations $\rho_n(X_i, Y)$ for variables X_i for $i = 1, \ldots, n$ with the response Y.
  Rank the variables in decreasing order of the sample distance correlations. Denote the ordered variables as $x_1, x_1, \ldots, x_n$. Start with $\mathbf{X_s} = \{\mathbf{x_1}\}$.
  for all i from 2 to n **do**
    Keep adding x_i to $\mathbf{X_s}$ if $\nu_n(\mathbf{X_s Y})$, the sample distance covariance, does not decrease. Stop otherwise.
  end for
return ($\mathbf{X_s}$)
end function

10. Diverse and Relevant Feature Selection

A methodological way of obtaining a solution for Problem III is by first running Kong-Wang-Wahba's All-relevant feature selection algorithm followed by running our proposed GreedyDiverseDCoV algorithm on the resulting solution of Kong-Wang-Wahba's algorithm. This would give a subset of the maximally separated diverse features that are also relevant with respect to the response. An alternate methodology would be to do the vice-versa of running our proposed GreedyDiverseDCoV algorithm first followed by running Kong-Wang-Wahba's algorithm on the resulting solution subset which is a union of the maximally separated subsets provided by our algorithm. This methodology of one before the other is analogous in principle to the forward selection or backward selection methods for variable(feature) selection. That said this methodology of running both the GreedyDiverseDCoV and Kong-Wang-Wahba's algorithms in series come with varied and useful theoretical guarantees as discussed in this paper and also deal with multiple objectives of diverse and relevant feature selection while also being completely model-free, free of distributional assumptions and being non-parametric.

11. Experiments

In this section we evaluate our above proposed combination of DiverseMinimalMaximDCoV Algorithm in Algorithm 1 for diverse selection applied on the subset returned by the relevant selection algorithm of Kong-Wang-Wahba in Algorithm 2. We compare this combination of diversity and relevancy encouraging feature selection with the mRMR Ensemble algorithm in [1] which also aims to select relevant and non-redundant (diverse) features.

11.1. Datasets used in experiments:

These are the three real-life datasets on which we evaluated the combination of Algorithm 1 on the results of Algorithm 2 and compared it with the mRMR Ensemble algorithm:

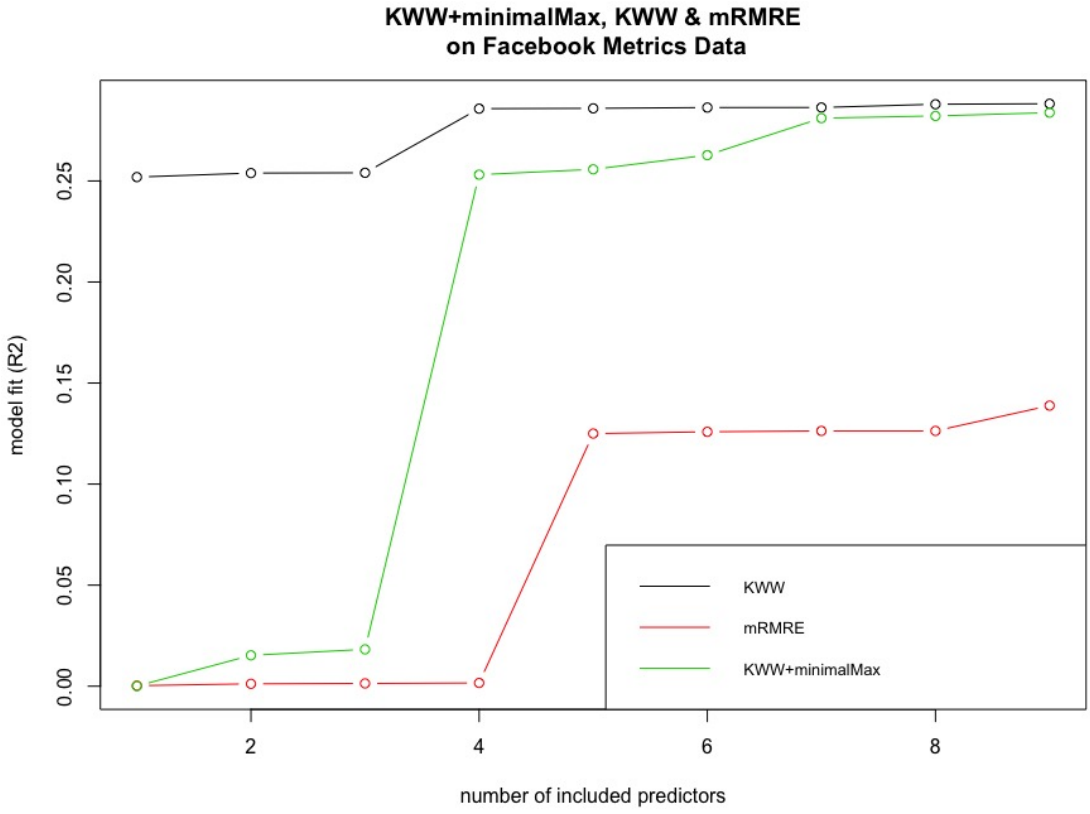

Figure 1: Results on UCI's Facebook's comment volume prediction dataset, https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset

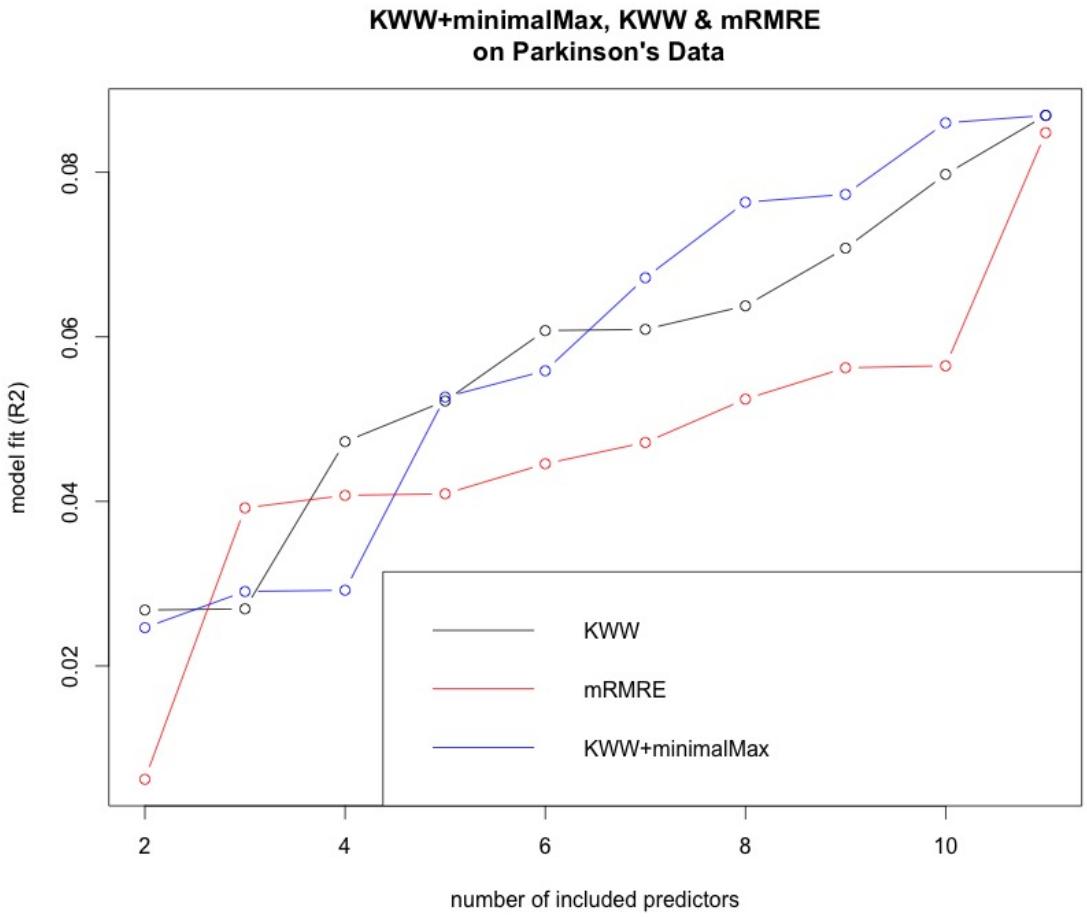

Figure 2: Results on UCI's Parkinson Speech Dataset with Multiple Types of Sound Recordings, https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with++Multiple+Types+of+Sound+Recordings

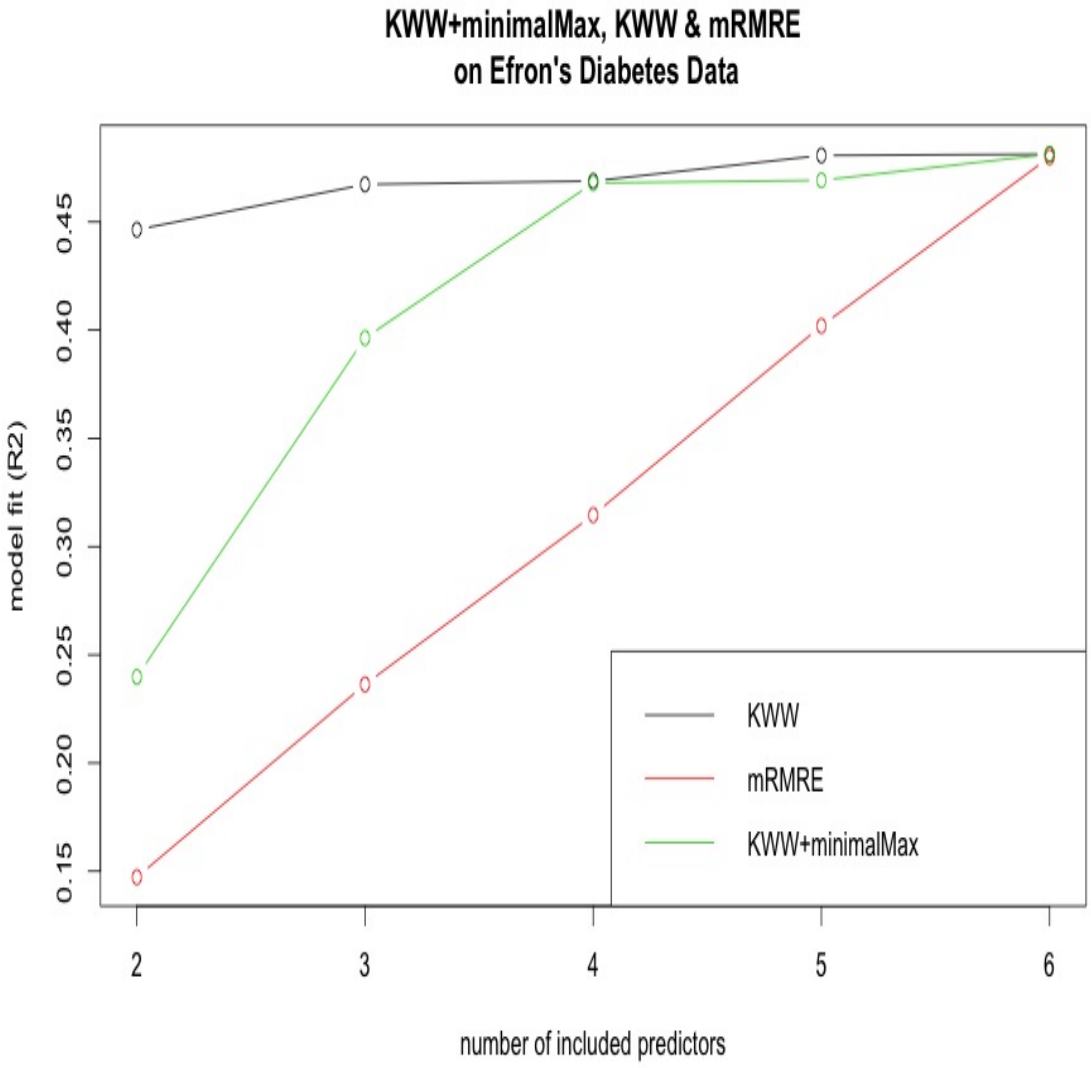

Figure 3: Results on Diabetes data of 442 patients from Efron et al. 2004. Least angle regression, Annals of Statistics, 32:407-499, http://artax.karlin.mff.cuni.cz/r-help/library/care/html/efron2004.html

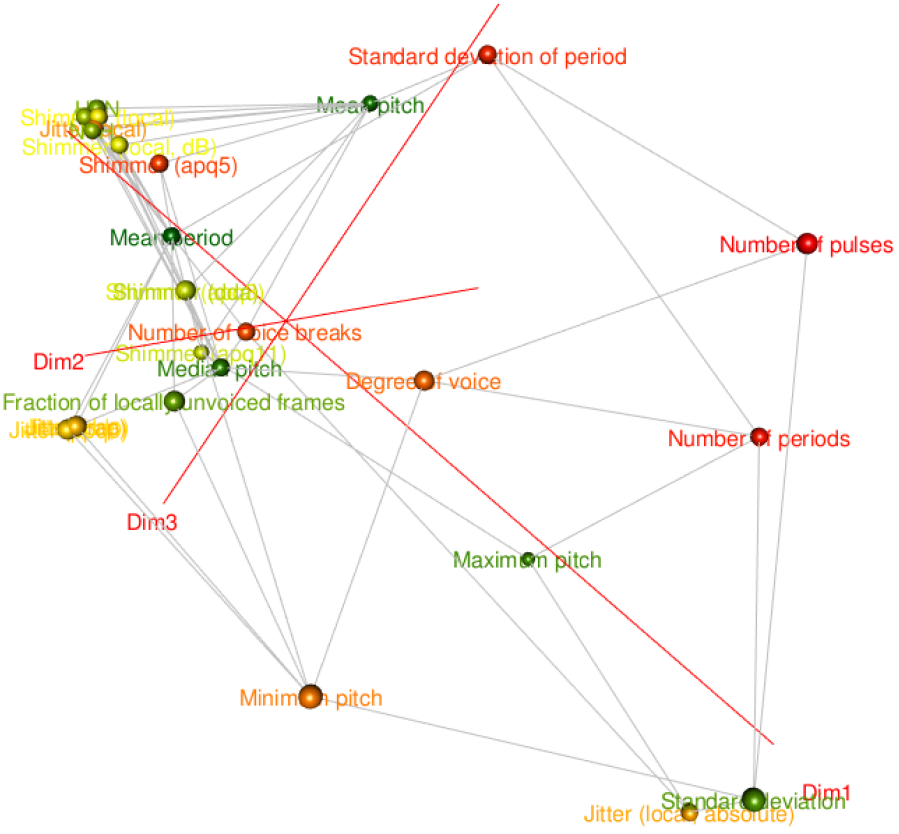

Figure 4: ISOMAP Embedding:

1. **UCI's Facebook's comment volume prediction dataset:**
   The goal associated with this dataset is to be able to predict the volume of comments on Facebook using various input metrics.

2. **UCI's Parkinson Speech Dataset with Multiple Types of Sound Recordings:**
   We aimed to use speech data from Parkinson's patients with varying levels of severity and non-patients inorder to predict the UPDRS score, a score that is widely used in the medical fraternity to gauge the severity of Parkinson's in the subject under investigation.

3. **Efron's Diabetes data:**
   This dataset consists of ten baseline variables of age, sex, body mass index, average blood pressure, and six blood serum measurements that were obtained for each of 442 diabetes patients, along with a response of interest, a quantitative measure of disease progression one year after baseline that was also collected. The goal associated is to be able to build a model that predicts this measure of disease progression.

We present the results of this algorithmic comparison evaluated by the R-Squared Error metric upon fitting linear regression models on these three datasets in Figures 1, 2 and 3. As seen, our approach clearly outperformed the mRMR Ensemble model. In the case of our proposed Algorithm 1 we took an iterative approach of obtaining the minimal maximizer subsets and then noting them and removing them to regenerate minimal maximizers from remaining set of features. We continued this process till the end or till enough number of features were generated. We do note that the quality of the first iterate of minimal maximizers with regards to optimization of our proposed objective will be higher than subsequent subsets of minimal maximizers, but this is amongst the best one could do in our setting in order to generate an entire ordering of features from being most diverse with regards to our objective to the least diverse.

In addition to this, it is interesting to analyze the gap between KWW+minimalMax line and KWW lines on the plot as it tells us how much the optimally relevant covariate selection of KWW matches with optimally diverse feature selection of our algorithm for any given dataset. So it tells us about the trade-off between relevancy and diversity of the covariates in any dataset like in a pareto frontier. Sometimes the relevancy maximizing and diversity maximizing subsets can intersect more and sometimes less based on the quality of the dataset in balancing these two criteria. Therefore this gap if quantified (say for example by integrating the difference between these lines) could be a good measure of evaluating the quality of any given dataset with respect to the relevancy-diversity tradeoff curve. This is just a direction we are pointing at and is not the main focus of our current paper.

In addition to this evaluation, we also performed a qualitative (and approximate) experiment to visually validate the diversity encouraging property of our theory. We did this by applying ISOMAP, a popular manifold learning technique on a matrix of pair-wise distance correlations between all pairs of features. This basically tries to generate a 2 dimensional Euclidean embedding like representation of of the Parkinson's dataset. This was presented in Figure 4 where we clearly were able to find the minimal maximizers produced by our

algorithm to be farther from the rest of the features (as colored in red). We actually color coded the points from red to green in the order generated by our proposed algorithm 1. Therefore we would expect the red features to be more diverse than the green. Although this figure is an approximation of the behavior of features with regards to diversity, it still somewhat matches visually with exact solution of our formulation.

In addition to comparisons with linear regression models on features selected by our approach and mRMR Ensemble, we also computed the 5 fold Cross- validated Mean Squared Error (MSE error) in predicting UPDRS scores with the Parkinson's dataset upon applying the random forest method of regression. Our combine approach of Algorithm 2 + Algorithm 1 produces a lower MSE of 148.84 vs mRMRe which obtained 154.39 MSE.

12. An efficient pre-processing routine: The effect of scaling and centering on combinatorics of $\nu_n(\cdot)$ and $\rho_n(\cdot)$:

We finally present an enumerative computational experiment we did to show that centering and scaling the data prior to applying our algorithms would lead to much better results as the distance covariances match up much better with distance correlations upon centering and scaling the data. This leads to the optimization of our proposed functions of distance covariance to auxiliarily mimic the optimization of our objective with distance correlation in the place of distance covariance. That is desirable as distance correlation is a normalized version of distance covariance.

As part of these empirical enumerative experiments, we collected various popular real-life regression and classification datasets from the well known University of California-Irvine Machine Learning Repository (UCI-ML) and enumerated the entire power set of possible combinations of their features (covariates) $2^{\mathbf{X}}$. We then computed the distance correlations between each subset belonging to the power set and the response (or class-label) variable $\mathbf{Y}$. We denote these distance correlations by $\rho_{\mathbf{E}}$. We also computed the distance covariances between each subset belonging to the power set and the response (or class-label) variable $\mathbf{Y}$. We denote these distance covariances by $\nu_{\mathbf{E}}$ in the same arbitrary order of subsets used when computing $\rho_{\mathbf{E}}$. Now with this set of paired measurements of $\rho_{\mathbf{E}}, \nu_{\mathbf{E}}$ available across the entire power set of combinations of features we computed the distance correlation of $\rho_{\mathbf{E}}, \nu_{\mathbf{E}}$ which we denote by $\rho(\rho_{\mathbf{E}}, \nu_{\mathbf{E}})$ to see if combinatorially optimizing distance covariance over the power set is a good proxy (surrogate) for combinatorially optimizing distance correlation. The distance correlation $\rho(\rho_{\mathbf{E}}, \nu_{\mathbf{E}})$ happened to be very high in almost all cases and very close to the theoretical upper-bound of 1 which indicates a strong statistical dependence between $\rho_{\mathbf{E}}$ and $\nu_{\mathbf{E}}$, thereby directly pointing out to the fact that combinatorially optimizing $\nu_{\mathbf{S}}, S \subseteq 2^{\mathbf{X}}$ is a great proxy for combinatorially optimizing $\rho_{\mathbf{S}}$ over the power-set. We would also like to mention that the values were close to one in the case when the covariates(features or variables) were centered and scaled; an operation that is a widely accepted pre-processing for regression or classification modeling. We were motivated to contrast the highly-encouraging results produced after centering and scaling with respect to not performing a centering and scaling because of the fact that the sample distance correlation is a function of sample distance covariances and for a fixed response variable $\mathbf{Y}$, the numerator of sample distance

correlation in equation 2 is dependent on both $\mathbf{X}$ and Y, while the denominator is only a function of $\mathbf{X}$ for a fixed response $\mathbf{Y}$. Thereby, the contribution of $||\ \mathbf{X}\ ||$ on $\rho_n(\mathbf{X}, \mathbf{Y})$ when $\rho_n(\mathbf{X}, \mathbf{Y})$ can be reduced by scaling and centering the data prior to computing the distance covariance. This can be further motivated by the following identity that was proved in [8]

$$\nu_n(\mathbf{X}, \mathbf{Y}) = \mathbf{Tr}\left(\mathbf{X^T L_Y X}\right) = \frac{1}{2}\sum_{i,j=1}^{n} [\widehat{\mathbf{E}}_\mathbf{Y}]_{i,j}[\mathbf{E_X}]_{i,j}.$$

where $\widehat{\mathbf{E}}_\mathbf{Y}$ is the double-centered Euclidean distance matrix formed with the rows of $\mathbf{Y}$ being the points for computing the pair-wise distances on and $\mathbf{E_Y}$ is the standard (without double-centering) Euclidean distance matrix of the rows of $\mathbf{X}$. This gives us that when $\mathbf{X} = \mathbf{Y}$, the denominator of distance correlation is solely a function of $\mathbf{X}$ that can be standardized across $\mathbf{S} \in \mathbf{2^X}$ by scaling and centering the values in $\mathbf{S}$.

All these results and comparisons of our enumerative experiment on the UCI-ML datasets are presented in Table 1 below.

Dataset	**Dimensionality**	$\mathbf{\lvert 2^X \rvert - 1}$	$\rho(\rho_\mathbf{E}, \nu_\mathbf{E})$ **without centering & scaling**	$\rho(\rho_\mathbf{E}, \nu_\mathbf{E})$ **with centering & scaling**
Airfoil Self-Noise	1503 by 5	31	0.896	0.999
Abalone	4177 by 8	255	0.422	0.693
Banknote Authentication	1372 by 4	15	0.938	0.993
Concrete Compressive Strength	1030 by 8	255	0.961	0.965
Protein Localization Sites of E.coli	336 by 7	127	0.891	0.966
Forest Fires	517 by 12	4095	0.841	0.941
Yacht Hydrodynamics	308 by 6	63	0.896	0.999

Table 1: A enumerative experiment with distance correlation and distance covariances over the power set

13. Conclusion:

We showed that our proposed Algorithm 1 gives exact solutions that are minimal-maximizers of our diversity encouraging objective. Similarly Algorithm 2 gives optimal solutions for a relevancy encouraging objective function. Now the quality of a solution subset that has a mixture of both properties of relevancy and diversity is dependent on pareto like trade-offs used in choosing the extent of diversity or relevancy one is willing to part away with unlike in highly optimal situations where the optimal solution of Algorithm 1 coincides with the optimal solution of Algorithm 2. That particular case would imply that the quality of the dataset being used for regression or classification is pretty optimal with regards to the relevancy-diversity tradeoff.

Bibliography:

[1] 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, pp. 1226-38, (2005).

[2] A. Das, A. Dasgupta and R.Kumar, *Selecting Diverse Features via Spectral Regularization*, Proceedings of Neural Information Processing Systems (NIPS), (2012).

[3] G. J. Szekely, M. L. Rizzo and N. K. Bakirov, *Measuring and Testing Dependence by Correlation of Distances*, The Annals of Statistics, 35(6), pp.2769-2794, (2007).

[4] Y. Kempner, B. Mirkin and I.Muchnik, *Monotone linkage clustering and quasi-concave set functions*, Applied Mathematics Letters, 10, pp.19-24, (1997).

[5] Michael R. Kosorok, *Correction: Discussion of Brownian distance covariance*, (2010), Annals of Applied Statistics, Volume 7, Number 2, pg. 1247, (2013).

[6] Xiaoming Huo and Gabor J. Szekely, *Fast Computing for Distance Covariance*, Technometrics, Volume 58, Issue 4, pg. 435-447, (2016).

[7] Jing Kong, Sijian Wang and Grace Wahba, *Using distance covariance for improved variable selection with application to learning genetic risk models.*, Statistics in Medicine, Volume 34(10), (2015), pg. 1708-20.

[8] Praneeth Vepakomma, Chetan Tonde, Ahmed Elgammal, *Supervised Dimensionality Reduction via Distance Correlation Maximization.*, ArXiv, (2016).

[9] J. Mullat, *Extremal subsystems of monotone systems: I, II*, Automation and Remote Control Volume 37,(1976),pg. 758-766; 1286–1294.