2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Max-Dependence Regression

Pouria Fewzee, Ali-Akbar Samadani, Dana Kulić, Fakhri Karray
University of Waterloo, Waterloo, ON, Canada N2L 3G1

***Abstract*—This work proposes an approach for solving the linear regression problem by maximizing the dependence between prediction values and the response variable. The proposed algorithm uses the Hilbert-Schmidt independence criterion as a generic measure of dependence and can be used to maximize both nonlinear and linear dependencies. The algorithm is important in applications such as continuous analysis of affective speech, where linear dependence, or correlation, is commonly set as the measure of goodness of fit. The applicability of the proposed algorithm is verified using two synthetic, one affective speech, and one affective bodily posture datasets. Experimental results show that the proposed algorithm outperforms support vector regression (SVR) in 84% (264/314) of studied cases, and is noticeably faster than SVR, as an order of 25, on average.**

I. Introduction

Regression analysis considers the problem of parameter estimation for a model with continuous response variables and a set of explanatory variables. The most commonly used model for regression is the linear model, described by $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$, where y is the response variable, $x_1 \ldots x_p$ are the explanatory variables, β_0 is the bias term, and $\beta_1 \ldots \beta_p$ are the linear coefficients. A common approach for estimating the β coefficients is to solve the optimization problem that minimizes the squared prediction error via shrinkage methods [1]. These methods add a regularization term to the cost function to penalize those explanatory variables that, in relative terms, do not contribute to lowering of the prediction error. A popular shrinkage algorithm is support vector regression (SVR) [1], [2]. SVR minimizes a regularized squared prediction error that is insensitive to errors smaller than a certain amount [3].

However, there are situations where in addition to minimizing the prediction error, the strength of association between explanatory and response variables is important (e.g., continuous analysis of affective speech [4], [5]). In such cases, maximizing a dependence measure between the explanatory and response variables is favored over minimizing the prediction error. Furthermore, the linear model, may not be the best choice when the response and explanatory variables are related in a nonlinear fashion, or when the linearity assumption does not result in an accurate enough approximation. To address this shortcoming, various solutions such as generalized linear models [6] and kernel methods [3] are employed.

In this work, we propose a novel regression approach that makes predictions based on a mapping of explanatory variables that maximizes statistical dependencies with the response variable. The maximization identifies a hypersurface along which minimizing the prediction error preserves the maximum dependencies between the mapped explanatory variables and the response variable; resulting in a prediction that is maximally correlated with the response variable and has the minimum error. This is in contrast to conventional linear regression approaches, where prediction error is minimized. The conventional approach does not guarantee maximum correlation between the predictions and response variables.

In particular, we distinguish between linear and nonlinear dependencies by using the Hilbert-Schmidt independence criterion (HSIC), a generic statistical dependence measure, and propose a solution for the regression problem in two stages: 1) extract a set of orthogonal transformations of explanatory variables that maximizes the nonlinear dependency with the response variable, and 2) construct a linear transformation over the mapped explanatory variables that maximizes the linear dependence between these variables and the response variable. HSIC has been previously used for dimensionality reduction [7], [8].

The performance of the proposed approach is evaluated and compared with the state-of-the-art SVR using synthetic datasets. Synthetic datasets enable examining the regression performance at different levels of nonlinearity, noise, and sample size. Furthermore, to validate the efficacy of the proposed approach for real-life applications, we apply it to predict affective dimensions for affective speech (VAM [9]) and affective posture (UCLIC [10]) datasets, and compare the results with those of SVR.

This paper is organized as follows: Section II describes the proposed regression approach and Section III presents the experimental procedure. Experimental results are presented in Section IV and discussed in Section V. We close the paper by conclusions and directions for future work in Section VI.

II. Methodology

Given a set of explanatory variables $\mathbf{x} \in \mathcal{X}$ ($\mathcal{X} \subset \mathbb{R}^p$) and a response variable $y \in \mathcal{Y}$ ($\mathcal{Y} \subset \mathbb{R}$), the objective is to find a dependence-maximizing linear mapping of $\mathcal{X}$ onto $\mathcal{Y}$. This can be formulated as the following optimization problem:

$$\underset{\beta}{\operatorname{argmax}} \quad \text{Dependence}(\mathbf{y}, \mathbf{X}\beta) \tag{1}$$

where $\mathbf{y}$ is an $N \times 1$ vector, $\mathbf{X}$ an $N \times p$ matrix, and β a $p \times 1$ vector, with N and p being the number of instances and the number of explanatory variables, respectively. We assume that the explanatory and response variables are standardized, i.e., each variable is normally distributed with a zero mean and standard deviation of one.

First we solve for the maximum correlation solution, that is linear dependence, and then we extend to the general notion of dependence using the Hilbert-Schmidt independence criterion (HSIC). There we get a series of vectors that are highly dependent on the response variable and are linearly independent among themselves. Therefore, to obtain the max-dependence solution, we use the solution obtained by maximizing correlation. In the following, we use lower and uppercase letters to denote scalars, lowercase bold-face to denote vectors, and uppercase bold-face to denote matrices. Moreover, we follow the convention of using Greek letters for parameters, and Latin letters for data.

A. Pearson Correlation Coefficient

We start by considering the linear dependence criterion, i.e., the Pearson's correlation coefficient.

$$r(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\sigma_{\mathbf{y}\hat{\mathbf{y}}}}{\sigma_{\mathbf{y}}\sigma_{\hat{\mathbf{y}}}}, \tag{2}$$

where $\sigma_{\mathbf{y}}$ represents the standard deviation of the variable $\mathbf{y}$, and $\sigma_{\mathbf{y}\hat{\mathbf{y}}}$ denotes the covariance of the two variables $\mathbf{y}$ and $\hat{\mathbf{y}}$. Given that we are seeking the linear mapping β that maximizes $r(\mathbf{y}, \hat{\mathbf{y}} = \mathbf{X}\beta)$, we can formulate the optimization problem as follows:

$$\underset{\beta}{\text{argmax}} \quad \frac{\sigma_{\mathbf{y}\mathbf{X}\beta}}{\sigma_{\mathbf{y}}\sigma_{\mathbf{X}\beta}}. \tag{3}$$

We can disregard the first term in the denominator, i.e., $\sigma_{\mathbf{y}}$, since it is independent of β. We force the standard deviation of the other term in the denominator to be one, since it only affects the optimal β by a scaling factor. We then have:

$$\begin{aligned} &\underset{\beta}{\text{argmax}} && \sigma_{\mathbf{y}\mathbf{X}\beta}, \\ &\text{subject to} && \sigma_{\mathbf{X}\beta} = 1. \end{aligned} \tag{4}$$

Using Lagrange multipliers and replacing the covariance and standard deviation with their estimates, we have

$$\frac{1}{N-1}\mathbf{y}^\top\mathbf{X}\beta + \lambda(1 - \frac{1}{N-1}\beta^\top\mathbf{X}^\top\mathbf{X}\beta) = 0. \tag{5}$$

Then, by taking the derivative with respect to the control parameter β, we have

$$\mathbf{y}^\top\mathbf{X} - 2\lambda\beta^\top\mathbf{X}^\top\mathbf{X} = 0, \tag{6}$$

which through some algebraic manipulation leads us to the solution of the optimization problem:

$$\beta_{\text{CC}} \propto (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{7}$$

This solution is identical to the solution of the ordinary least squares (OLS) estimator. That is to say, the OLS estimate maximizes the Pearson's correlation coefficient, which could be advantageous due to the well-behaved properties of the OLS, and moreover the variety of methodologies that are developed around ordinary least squares [1].

Despite the upsides of OLS, it is unable to account for a more general sense of dependence. However, if one could capture those dependencies in the form of a number of linearly independent components, then OLS built on those components would be a valid solution to the problem. To address shortcoming of OLS, we consider another notion of independence, the Hilbert-Schmidt independence criterion (HSIC). The promise of HSIC is that it defines dependence in the general sense, since it is established on the kernel spaces of the explanatory and response variables.

B. Hilbert Schmidt Independence Criterion

Assuming $\mathcal{F}$ and $\mathcal{G}$ to be two separable reproducing kernel Hilbert spaces [11] and $\mathcal{Z} := \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\} \subseteq \mathcal{X} \times \mathcal{Y}$, HSIC is defined as follows:

$$\begin{aligned} \text{HSIC}(p_{x,y}, \mathcal{F}, \mathcal{G}) = \mathbf{E}_{x,x',y,y'}[k(x,x')l(y,y')] \\ +\mathbf{E}_{x,x'}[k(x,x')]\mathbf{E}_{y,y'}[l(y,y')] \\ -2\mathbf{E}_{x,y}[\mathbf{E}_{x'}[k(x,x')]E_{y'}[l(y,y')]], \end{aligned} \tag{8}$$

where pairs of (x, y) are drawn from the joint probability distribution of $\mathcal{X}$ and $\mathcal{Y}$ represented by $p_{x,y}$. $\mathbf{E}$ denotes the expectation operator. To enable approximation given a finite number of samples, the empirical estimate of HSIC [12] is introduced as follows:

$$\text{HSIC}(\mathcal{Z}, \mathcal{F}, \mathcal{G}) = (N-1)^{-2}\mathbf{tr}(\mathbf{KHLH}). \tag{9}$$

Where $\mathbf{K}$,$\mathbf{L}$,$\mathbf{H} \in \mathbb{R}^{N\times N}$, $K_{ij} := k(\mathbf{x}_i, \mathbf{x}_j)$, $L_{ij} := l(\mathbf{y}_i, \mathbf{y}_j)$, and $H := \mathbf{I} - N^{-1}\mathbf{e}\mathbf{e}^T$, where $\mathbf{e}$ is a vector of N ones. It can be shown that the HSIC of two independent variables is zero. Therefore, by assuming that $\mathbf{K}$ represents a kernel of the linear projection, that is $\mathbf{X}\beta$, and $\mathbf{L}$ a kernel of the response variable $\mathbf{y}$, what is of interest is the mapping β that maximizes $\mathbf{tr}(\mathbf{KHLH})$[8]. By further assuming that the two kernels are linear, i.e., $\mathbf{K} = \mathbf{X}\beta\beta^\top\mathbf{X}^\top$ and $\mathbf{L} = \mathbf{y}\mathbf{y}^\top$, we have:

$$\begin{aligned} \mathbf{tr}(\mathbf{KHLH}) &= \mathbf{tr}(\mathbf{X}\beta\beta^\top\mathbf{X}^\top\mathbf{H}\mathbf{y}\mathbf{y}^\top\mathbf{H}) \\ &= \mathbf{tr}(\beta^\top\mathbf{X}^\top\mathbf{H}\mathbf{y}\mathbf{y}^\top\mathbf{H}\mathbf{X}\beta) \end{aligned}$$

Hence, we are interested in the solution of the following optimization problem.

$$\begin{aligned} &\underset{\beta}{\text{argmax}} && \mathbf{tr}(\beta^\top\mathbf{Q}\beta), \\ &\text{subject to} && \beta^\top\beta = \mathbf{I}, \end{aligned} \tag{10}$$

where $\mathbf{Q} = \mathbf{X}^\top\mathbf{HLHX}$. The constraint is to make the optimization problem well defined, since in its absence, it is unbound. Through a set of algebraic manipulations, it can be shown that the solution to this optimization problem is the eigenvectors of $\mathbf{X}^\top\mathbf{HLHX}$ that correspond to the top eigenvalues.

If the kernels are linear, maximizing HSIC is equivalent to maximizing the Pearson's correlation coefficient. Extension to nonlinear kernels is straightforward [8].

C. Max-Dependence Regression

With the objective of maximizing the dependence between the response variable and the linear mapping of the explanatory variables, as a solution to the regression problem, we propose the following algorithm:

1. $\mathbf{Q} \leftarrow \mathbf{X}^\top\mathbf{HLHX}$
2. Let columns of $\mathbf{V}$ be the eigenvectors of $\mathbf{Q}$
3. $\beta_{\text{HSIC}} \leftarrow \mathbf{V}_{\mathcal{S}}$, where $\mathcal{S}$ represents the selected subset of $\mathbf{V}$ columns.
4. $\hat{\mathbf{X}} \leftarrow \mathbf{X}\beta_{\text{HSIC}}$
5. $\beta_{\text{CC}} \leftarrow (\hat{\mathbf{X}}^\top\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^\top\mathbf{y}$
6. $\beta_{\text{MDR}} \leftarrow \beta_{\text{HSIC}} \cdot \beta_{\text{CC}}$

Steps 1-3 encapsulate the required operations for extracting components that are maximally dependent on the response variable. At this stage nonlinear dependence is of interest. Steps 4-5 find a linear combination of the components from the previous stage that maximizes the overall correlation with the response variable. Finally, Step 6 aggregates the two stages.

III. Experimental Setup

The performance of MDR is evaluated and compared with SVR using synthetic and real datasets.

A. Synthetic Datasets

Synthetic datasets were used to enable assessment of the proposed approach at varying levels of non-linearity, noise, and size. Different synthetic datasets were tested. For the sake of brevity, we present results from two of these datasets.

1) **Dataset.1:** the regression model is defined as:

$$Y = X_1(\sin(2\pi f X_2) + 1) + \epsilon, \tag{11}$$

2) **Dataset.2:** the regression model is defined as:

$$Y = \mathrm{sinc}(2\pi f X_2) + \epsilon. \tag{12}$$

In these synthetic datasets, $X_1 \in [0, 1]$ and $X_2 \in [0, 1]$ are uniformly distributed random variables, f is the frequency, and ϵ is a normal additive noise $\epsilon \sim N(0, \sigma^2)$. σ^2 is the variance of the normal noise.

B. Affective datasets

To evaluate the performance of the proposed MDR approach with real datasets, it is applied on two affective datasets: 1) affective speech [9], 2) affective posture [10], and its performance is compared with that of SVR.

Using the dimensional representation approach, affective expressions are described as points in a continuum of a low-dimensional space [13]. For instance, the Circumplex model [14] represents affective expressions in a two dimensional space defined by arousal and valence. The arousal (activation) dimension represents the level of activation, mental alertness, and physical activity, whereas the valence dimension ranges from negative (unpleasant) to positive (pleasant). In this work, the dimensional representation of affective expressions is used.

1) Speech Dataset: VAM [9] is a spontaneous emotional speech dataset, collected from the recordings of the German talk show *Vera am Mittag*. A total of 47 subjects, 36 female and 11 male, take part in the recordings, and their ages range from 16 to 69, where 70% of the actors are 35 or younger. VAM-Audio includes two modules, VAM-Audio I and II, comprising about 50 minutes of recording in total. The division into two modules is based on the quality of the conveyed emotions. VAM-Audio I, classified as very good, contains 19 speakers, making 478 utterances. This is the part of the dataset that we use. On the other hand, VAM-Audio II, classified as good, contains 28 speakers and a total of 469 utterances. Various studies of emotional speech have been conducted based on VAM [9], [15], [16], [17], and [18], [19].

In VAM-Audio I, each sample is annotated by 17 observers on three affective dimensions: activation, dominance, and valence, and the median of observers' annotations for each dimension is used as the measure of that affective dimension (response variable) in our experiments. As a measure of agreement between the observers, the Cronbach's alpha (α) is computed for each of the dimensions (Table I).

TABLE I. Cronbach's alpha as a measure of agreements between the observers' ratings for the speech and posture data.

Speech			Posture			
Activation	Dominance	Valence	Arousal	Avoidance	Potency	Valence
0.97	0.94	0.85	0.88	0.52	0.54	0.81

We use the power spectrum filter banks of the signal, summarized by their statistics for the low level speech descriptor, to make for a vector of 595 features. We assume that these features collectively have a high correlation with affective dimensions [20], and form non-linear regression models of activation, dominance and valence as a function of the extracted feature vector.

2) Posture Dataset: The posture dataset contains 110 affective postures representing the most expressive frames in affective movements from the UCLIC dataset [10]. These movements are captured from thirteen demonstrators who freely expressed movements conveying anger, happiness, fear, and sadness, without any kinematic constraints on the movements. The most expressive frame of the movement was selected by its demonstrator. There are 32 markers attached to bodily landmarks and their 3D Cartesian coordinates are collected using a motion capture system; hence a total of 96 Cartesian coordinates for each posture. There are 25 sad, 21 happy, 40 fearful, and 24 angry postures in the posture dataset.

The affective postures were rated by 10 observers along arousal, valence, potency (dominance), and avoidance (avoid/attend to) dimensions using 7-point Likert scales. The Cronbach's alpha (α) is computed for each affective dimension to assess the strength of agreement between the observers (Table I).

For our analysis, the affective postures are centered horizontally and rotated to align their torso line (lateral waist markers) along the positive x-axis followed by minmax normalization of the postures' Cartesian coordinates. For each posture, the median of observers' ratings for an affective dimension is used as the measure of that affective dimension (response variable). We assume that postures are highly correlated with the affective dimensions and formulate non-linear regression models for each affective dimension as a function of posture.

C. Experimental procedure

The performance of the proposed MDR approach is evaluated and compared with that SVR using 10 repetitions of 10-fold cross validation (referred to as 10×10FCV, hereafter) on the synthetic and real datasets. 10-fold cross validation is used for its reliability in model selection and accuracy estimation [21], [22]. The same settings for the folds are used to test the performances of both approaches.

The synthetic datasets, as described by Equations 11 and 12, are generated using different combinations of 3 sample sizes (50, 100, 500), 10 frequencies (0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 128, 1024), and 5 noise ratio levels (0.0125, 0.025, 0.05, 0.5, 1), for a total of 150 cases. For the synthetic datasets, to equalize conditions under which MDR and SVR are compared, the linear implementation of both approaches is tested. It is clear that a kernelized implementation of SVR and MDR is more suited for cases where a high-level of nonlinearity is introduced in the synthetic datasets. However, implementing kernelized SVR and MDR requires selecting a suitable kernel and tuning its hyper-parameters, which adds to the number of conditions under which the two approaches are compared. While we recognize the importance of comparing the relative performance of the kernelized SVR and MDR with synthetic nonlinear datasets, for the sake of simplicity of analysis (number of comparison conditions) and to maintain an equal ground for comparing the two approaches, we only implement linear versions of the two approaches in the present work. An extended comparison based on kernelized SVR and MDR with synthetic nonlinear datasets is a future direction for this work.

For the affective datasets, the kernelized versions of MDR and SVR are used, where we have considered linear, radial basis, and polynomial kernels. The kernel types and their hyper-parameters for MDR and SVR and the SVR's slack parameter are selected to optimize Pearson's correlation coefficient in a cross validation test performed on the training set[1]. In the experiments with the synthetic datasets, the SVR slack parameter is selected in the same manner.

There are different sources of variation in the expression of affect, including person-specific and idiosyncratic variations. In order to test the generalization ability of the proposed approach to different subjects, leave-one-subject-out cross validation (LOSOCV) is also performed for the speech and posture datasets, each including 19 and 13 subjects, respectively. In each fold of LOSOCV, a subject is left out (testing subject) and the models are trained using the remaining subjects (training subjects).

Cross correlation (CC) and mean absolute percentage error (MAPE) are used for evaluation. Additionally, training and recall times are used to compare the computational complexity of the algorithms.

IV. Results

In this section, the results of the experimental evaluation are reported for the synthetic datasets in Section IV-A and for affective datasets in Sections IV-B and IV-C.

A. Synthetic Datasets

The results on the two synthetic datasets are shown in Fig. 1, where each point corresponds to an experiment with samples generated given a sample size, a frequency, and a noise ratio, and abscissa and ordinate of each point indicate the resulting correlation coefficient by SVR and MDR, respectively. We use the relative position of the points with respect to the identity (1:1) line to assess the relative performance of the two approaches in each scenario. Points that are on the top side of the line favor MDR over SVR, and points that are on the bottom side favor SVR over MDR. The further a point gets from the line, the more one approach is favored over the other.

1) Dataset.1: Fig. 1(a) presents the results of the first synthetic dataset (Equation 11). According to this figure, in more than 95% of cases (143/150), MDR produces higher correlation than SVR. For 50 samples, MDR shows better performance than SVR in all cases. By increasing the sample size, we see an evident shift towards the identity line, and despite the better performance of MDR in the sample size of 500, points are very close to the identity line. The sum of distances of the points to the identity line for the sample sizes of 50, 100, and 500, are 4.03, 3.13, and 1.12. For this dataset, we could say that MDR shows better performance compared to SVR when few data points are available.

Increasing the frequency and/or noise ratio, decreases the overall performance of both methods. This is expected, given that these two parameters contribute to nonlinearity and unpredictability of data, respectively. However, the degree to which the two methods are affected by these changes is different. Fig. 2 shows the relative trend of changes of correlation with respect to sample size, frequency, and noise ratio. The ordinate of these figures indicates the percentage of cases where MDR results in a higher correlation than SVR. As the frequency or noise ratio increase, MDR's performance monotonously becomes better than that of SVR.

Average training and recall times are 106.4 and 0.6 milliseconds for MDR, and 2700.5 and 0.8 milliseconds for SVR. That is, MDR is more than 25 times faster than SVR in terms of training time, with similar recall time.

2) Dataset.2: Fig. 1(b) presents the results of the second synthetic dataset (Equation 12). According to this figure, in more than 75% of cases (113/150), MDR results in a higher correlation than SVR. In terms of sample size, a similar trend to the first synthetic dataset is observed. For the lowest sample size, MDR outperforms SVR, with increased sample size, they tend to show more similar results, still in favor of MDR for the higher sample sizes. The sum of distances of the points to the identity line for the samples sizes of 50, 100, and 500, are 4.47, 2.49, and 0.54.

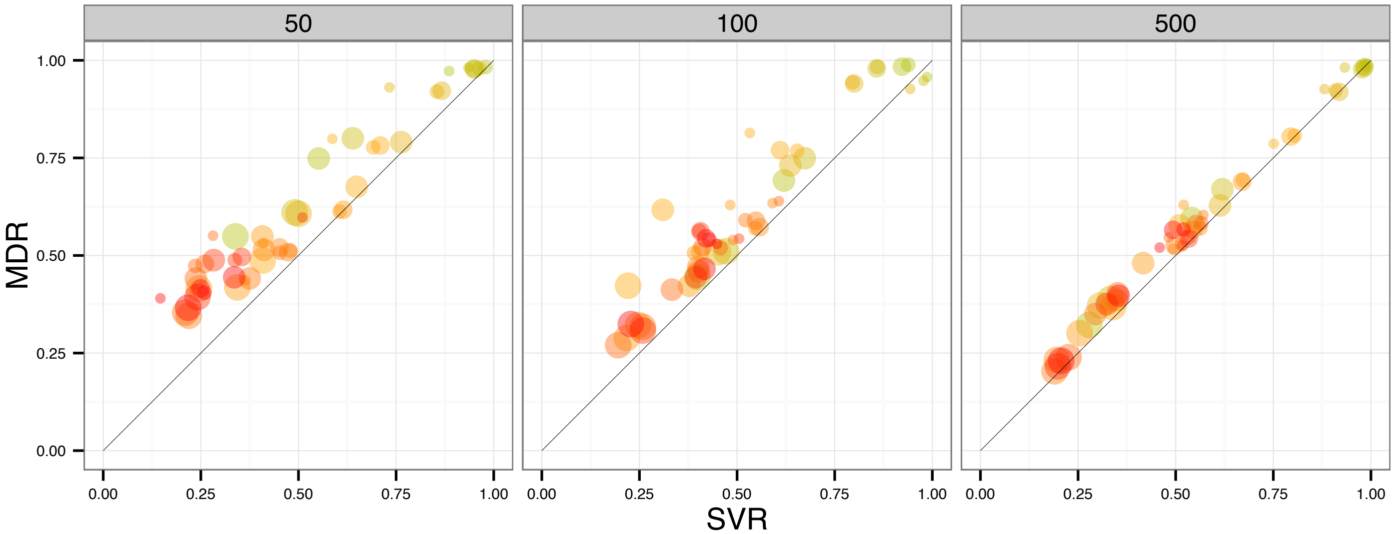
For this dataset, the trend does not seem to be as smooth (Fig. 2), however, in this case too the average correlation for MDR is higher than that of SVR. Despite the performance deterioration with the increase in the frequency and noise ratio, MDR results in a higher correlation than that of SVR. The only exception is the results at $f = 1024$, where SVR demonstrates a slightly better performance (Fig. 2). However, the correlation coefficients at this frequency in all cases fall below 30% for both approaches, hence, no conclusion can be derived on the superiority of one approach over the other.

Average training and recall times are 108.3 and 0.6 milliseconds for MDR, and 1465.1 and 0.8 milliseconds for SVR. That is, SVR is more than 13 times slower than MDR in terms of training time, however, they are similarly fast in the recall phase.

B. Speech Dataset

Table II shows average CC($\pm$ std) and MAPE($\pm$ std) for the predicted affective dimensions obtained by 10$\times$10FCV. For

[1] In each fold of 10$\times$10FCV and LOSO, a separate 5-fold cross validation is performed using only the training set, and kernels (and their hyper-parameters) maximizing CC are selected to perform regression in that fold.



(a) Dataset.1

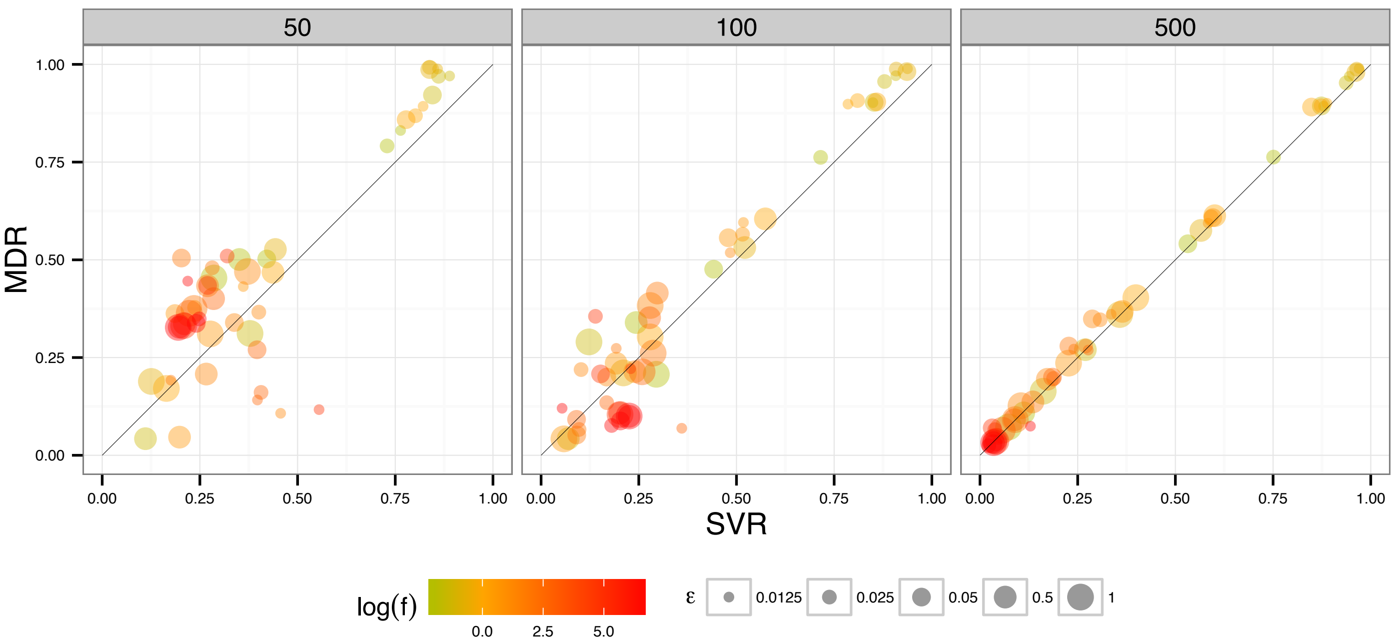


Fig. 1. Relative performance of MDR versus SVR for different combinations of sample size (50, 100, and 500), frequency (f), and noise ratio (ϵ). Points that are above (below) the line are those that favor MDR (SVR).

both SVR and MDR, the best results were obtained with the radial basis kernel. The high CC and low MAPE resulting from both approaches in the prediction of the activation and dominance dimensions show the high accordance of the predicted values with those perceived by the observers.

Unlike the activation and dominance dimensions, for which MDR and SVR perform equally well, the performance is significantly poorer for valence. This is in spite of the fact that the Cronbach's alpha for valence is in the *good* range, meaning that the agreement between the observers is relatively high (Table I). A possible explanation is that the observers' evaluation is based on both the audio and visual modalities, and that the two modalities are not equally effective in conveying different dimensions of affect. Since only the audio part of the dataset is used for regression, the low level of correlation in predicting valence might be due to the insufficiency of the explanatory variables.

To further examine the capability of the proposed approach in generalizing to unseen subjects, leave-one-subject-out cross validation experiments are conducted. The results of those experiments are shown in terms of the CC and MAPE in Table II. The trend here is very similar to that of the 10×10FCV, where both MDR and SVR show similar performance in predicting the activation and dominance dimensions of unseen subjects, and considerably lower performance in predicting the valence. Although the two approaches do not show a meaningful difference in predicting activation and dominance, the difference is noticeable for the valence.

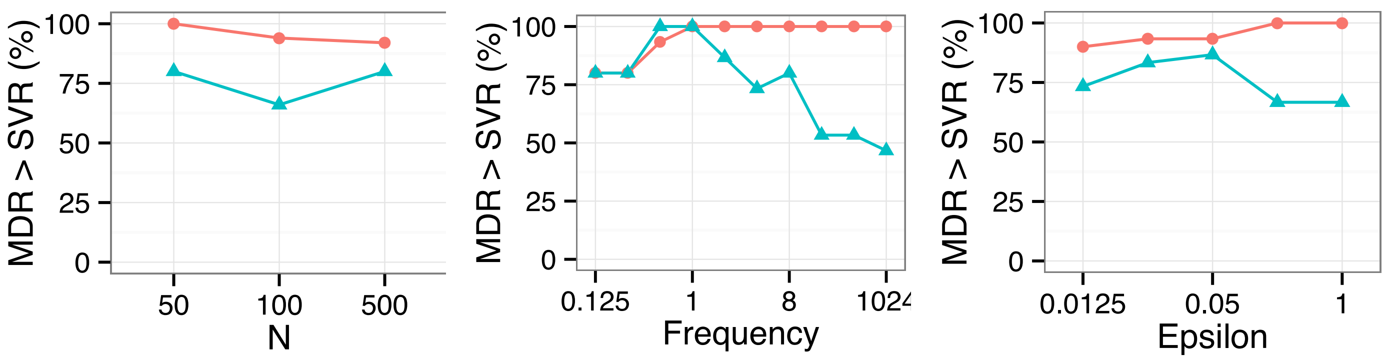


Fig. 2. Trends of changes of CC with respect to sample size, frequency, and noise ratio. Disk and triangle correspond to the dataset.1 and 2, respectively.

C. Posture Dataset

Table III shows average CC($\pm$ std) and MAPE($\pm$ std) for the predicted affective dimensions, obtained by 10$\times$10fold cross validation (10$\times$10FCV). Similar to the previous case, for both SVR and MDR, the best results were obtained with the radial basis kernel. The high CC and low MAPE of both approaches in the prediction of the arousal and valence dimensions show the high accordance of the predicted values with those perceived by the observers.

Although MDR outperforms SVR in predicting potency, both approaches perform poorly for potency and avoidance. This poor prediction could be due to the poor agreement between observers in rating these dimensions (Cronbach α's reported in Table I). The perception of these dimensions might be ambiguous from postural cues alone, and additional modalities might be needed to correctly evaluate avoidance and potency. Nevertheless, the proposed MDR performs similarly to SVR in modeling human perception of affective dimensions and predicting these dimensions from postures. Despite the performance similarities, MDR demonstrates lower variations in predicting the affective dimensions as compared with SVR across the 10$\times$10FCV folds (Table III).

To further evaluate the generalization ability of the MDR model in predicting affective dimensions from person-specific postures, LOSOCV was conducted. The resulting CC and MAPE are shown in Table III[2].

Fig. 3 shows average ($\pm$ std) training and recall times for MDR and SVR across 10$\times$10FCV folds. MDR training and recall times are significantly shorter than those of the SVR.

V. Discussion

The experiments with the synthetic datasets were designed to evaluate the relative performance of MDR and SVR at varying levels of non-linearity, unpredictability, and sample size. Non-linearity and unpredictability were introduced by varying frequency and noise ratio, respectively. Based on the results reported in Section IV-A, we can make the following hypotheses: 1) MDR outperforms SVR when few samples are available and the two approachs perform more similarly as the sample size increases, 2) SVR performs better than MDR at smaller noise ratios; at higher noise ratios MDR outperforms SVR, 3) the performance of both approaches deteriorates as the frequency (viz. nonlinearity) increases. The third hypothesis is weak due to the lack of experiments with kernelized MDR and SVR for the nonlinear datasets in the present work.

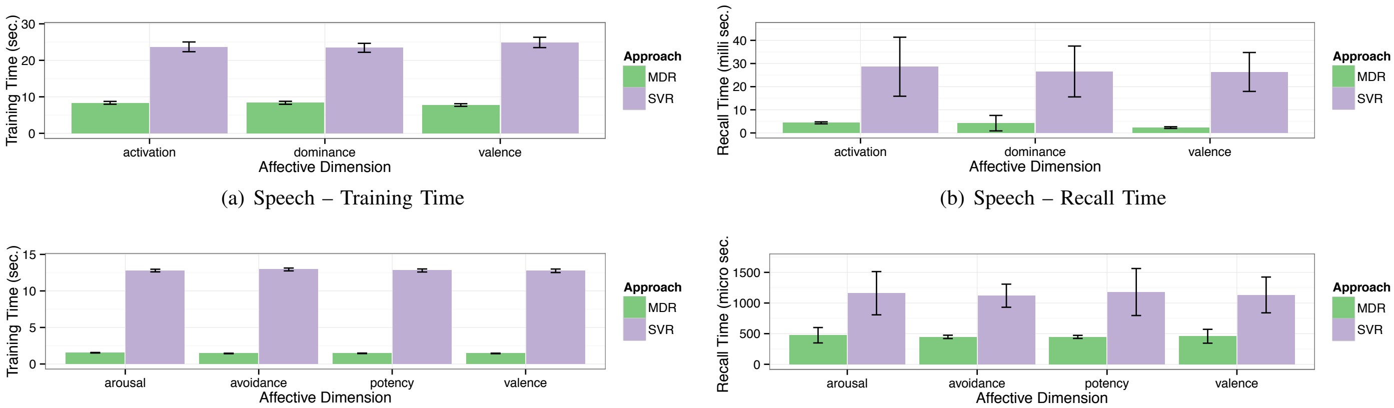
The first hypothesis also holds when comparing the regression performance for the speech and posture datasets with one another. The VAM Audio-I has 478 samples and the posture dataset has 110 samples. MDR's crossvalidated CC's are higher than those of SVR in the posture dataset (smaller sample size), while crossvalidated CC's for both approaches are close for the speech dataset (larger sample size).

To further evaluate the first hypothesis, additional experiments were conducted using the VAM dataset. A 10$\times$10 FCV was conducted using 20% of samples randomly selected from VAM Audio-I. The results from the experiment with a subset of VAM Audio-I dataset show that SVR outperforms MDR in terms of average cross-validated CC's (SVR $>$ MDR by: 2.83% in Activation, 7.06% in Dominance, and 2.51% in Valence). These results do not support the first hypothesis on the advantage of MDR over SVR for small sample sizes.

A similar poorer performance of MDR in comparison with SVR is also observed in the experiments with 50 samples of the synthetic dataset 2. As can been seen in Figure 1(a), in all such cases, the noise ratio is very low and as the noise ratio increases and sample size remains fixed, MDR surpasses SVR. A possible explanation is that there is an interaction effect between the noise ratio and sample size such that the effect of sample size varies at different levels of noise ratio. To test this hypothesis, we have rerun the 10$\times$10 FCV with 20% of samples randomly selected from VAM Audio-II where there is a lower agreement between observers on conveyed affective dimensions in comparison with VAM Audio-I; which in turn makes it more noisy than VAM Audio-I. VAM Audio-II contains 469 samples in total. On average, MDR performs better than SVR on the subset of VAM Audio-II dataset in terms of cross-validated CC's from 10$\times$10 FCV (MDR $>$ SVR by: 2.19% in Activation, 1.29% in Dominance, 1.66% in Valence).

Therefore, the relative performance of MDR and SVR on subsets of VAM Audio-I and VAM Audio-II shows that at a similar sample size, SVR outperforms MDR at lower

[2]Due to the imbalance in the number of postures from each subject, the computation of average performance metrics (CC and MAPE) is problematic. To overcome this problem, predictions from each fold of LOSOCV are concatenated and then an average CC and MAPE over all the predictions is computed; hence, no std for LOSOCV folds.



(a) Speech – Training Time

(b) Speech – Recall Time

(c) Posture – Training Time

(d) Posture – Recall Time

Fig. 3. Average training time ± std over folds of 10×10FCV. The error bars indicate standard deviation.

TABLE II. RESULTS ON THE AFFECTIVE SPEECH DATASET. (CC: CORRELATION COEFFICIENT, MAPE: MEAN ABSOLUTE PERCENTAGE ERROR)

	Activation		Dominance		Valence	
	CC	MAPE	CC	MAPE	CC	MAPE
	10×10-Fold Cross Validation					
SVR	82.08 ± 0.45	5.92 ± 0.04	76.10 ± 0.45	5.64 ± 0.05	46.72 ± 1.29	8.04 ± 0.14
MDR	82.15 ± 0.29	6.01 ± 0.05	77.36 ± 0.26	5.72 ± 0.03	43.43 ± 1.62	9.30 ± 0.11
	Leave-One-Subject-Out					
SVR	81.68	6.00	74.95	5.76	40.83	8.59
MDR	81.23	6.17	75.07	5.99	33.09	9.79

TABLE III. RESULTS ON THE AFFECTIVE POSTURE DATASET. (CC: CORRELATION COEFFICIENT, MAPE: MEAN ABSOLUTE PERCENTAGE ERROR)

	Arousal		Avoidance		Potency		Valence	
	CC	MAPE	CC	MAPE	CC	MAPE	CC	MAPE
	10×10-Fold Cross Validation							
SVR	73.44 ± 4.17	7.79 ± 0.46	41.68 ± 5.18	11.61 ± 0.51	20.39 ± 7.26	10.72 ± 0.48	63.00 ± 5.55	8.38 ± 0.47
MDR	77.10 ± 0.62	7.59 ± 0.12	42.08 ± 3.04	11.54 ± 0.27	37.38 ± 2.74	10.20 ± 0.26	59.66 ± 1.68	8.60 ± 0.24
	Leave-One-Subject-Out							
SVR	72.29	7.92	37.92	11.79	5.22	11.11	61.89	8.62
MDR	77.09	7.49	36.71	12.07	38.92	9.89	60.83	8.65

noise ratios (VAM Audio-I), while at higher noise ratios (VAM Audio-II), MDR outperforms SVR, which is congruent with the hypothesis on the interaction effect of noise ratio and sample size. These results also support the hypothesis regarding MDR's advantage at higher noise ratios (Hypothesis 2).

Another advantage of MDR over SVR is its computational efficiency. As shown in Section IV-A and Figure 3, MDR's training and recall times for the synthetic and real datasets are significantly shorter than those of SVR. As can be seen in Figure 3, MDR's recall time for each affective dimension is also more consistent across the 10×10FCV folds, whereas SVR's recall time shows a larger standard deviation. Furthermore, there are higher variations in SVR recall time across the affective dimensions, whereas MDR's recall time is consistently short (Figure 3).

Another important observation is that by decreasing the number of explanatory variables from two (synthetic dataset 1) to one (synthetic dataset 2), the average training time of SVR is almost halved (from 2700.5 ms in dataset 1 to 1465.1 ms in dataset 2), whereas MDR's training time did not meaningfully change (106.4 ms in dataset 1 and 108.3 ms in dataset 2). The importance of this difference could be even more evident in cases where the dimensionality of the feature space is large.

VI. CONCLUSION

In this work, we presented a new regression approach, max-dependence regression (MDR), that aims to make predictions that are maximally correlated with the response variable. The proposed approach exploits HSIC measure and identifies a mapping of explanatory variables that are maximally dependant to the response variable. The mapped explanatory variables are then linearly combined to produce the optimal prediction for the response variable. As a result, MDR maximizes both linear and nonlinear dependencies between the response and explanatory variables.

The proposed MDR approach was evaluated using synthetic and real datasets and its performance was compared with those of SVR. The synthetic datasets allowed us to evaluate the performance of the proposed approach under varying levels of noise, nonlinearity, and sample size in comparison with an state-of-the-art regression approach, SVR. Two synthetic datasets were examined and the prediction performance of MDR was found to surpass SVR's in majority of the cases.

In particular, MDR demonstrates better performance than SVR with lower samples sizes, with the two approaches demonstrating a similar performance as the sample size increases. In terms of unpredictability and nonlinearity, MDR shows a better performance than SVR in most cases where more unpredictability and nonlinearity were introduced to the datasets by increasing the noise ratio ans/or varying the frequency, respectively.

MDR was also applied to two affective datasets and its performance compared to that of SVR using two cross validation tests: 10×10 FCV and LOSOCV. Results show a close performance of the two regression approaches.

One clear advantage of MDR is its computational efficiency. MDR's training and recall times were shown to be significantly shorter than those of SVR in the experiments on both the synthetic and real datasets. Therefore, the proposed MDR performs similarly to (and in many cases surpasses) the state-of-the-art SVR, but with significantly shorter training and recall time.

Further experiments with additional datasets will be conducted in the future to examine the performance of MDR. In particular, the hypothesized main and interaction effects of sample size, noise ratio, and frequency will be further explored in the future. Moreover, experiments on the relative performance of kernelized MDR and SVR at varying levels of nonlinearity is a future direction for this work. Another direction for the future work is to extend MDR to sparsify the regression coefficients to include only the explanatory variables most relevant to the response variable. For this, synthetic datasets with varying number of features will be used where the response variable depends only on a subset of the variables. The importance of sparsification becomes evident in applications such as affective movement analysis where there are different sources of variations (e.g., interpersonal and kinematics) among which only a few are salient to affective expressions [23], [13].

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., ser. Springer series in statistics. Springer, 2008.

[2] C. M. Bishop, *Pattern recognition and machine learning*, 8th ed. Springer, 2009.

[3] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed., ser. Information Science and Statistics. Springer, 2000.

[4] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012 – the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, ser. ICMI '12. ACM, 2012, pp. 449–456.

[5] P. Fewzee and F. Karray, "Continuous emotion recognition: Another look at the regression problem," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 197 – 202.

[6] A. Dobson, *An Introduction to Generalized Linear Models*, ser. Texts in Statistical Science. Chapman & Hall/CRC, 2002.

[7] Y. Zhang and Z.-H. Zhou, "Multilabel dimensionality reduction via dependence maximization," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 4, no. 3, p. 14, 2010.

[8] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Z. Jahromi, "Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds," *Pattern Recognition*, vol. 44, no. 7, pp. 1357 – 1371, 2011.

[9] I. Kanluan, M. Grimm, and K. Kroschel, "Audio-visual emotion recognition using an emotion space concept," *Signal Processing*, 2008.

[10] A. Kleinsmith, P. De Silva, and N. Bianchi-Berthouze, "Cross-cultural differences in recognizing affect from body posture," *Interacting with Computers*, vol. 18, no. 6, pp. 1371–1389, 2006.

[11] M. Hein and O. Bousquet, "Kernels, associated structures and generalizations," Max Planck Institute for Biological Cybernetics, Tech. Rep. 127, 2004.

[12] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *Algorithmic Learning Theory*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2005, vol. 3734, pp. 63–77.

[13] M. Karg, A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, and D. Kulić, "Body movements for affective expression: A survey of automatic recognition and generation," *IEEE Trans. Affect. Comp.*, vol. 4, no. 4, pp. 341–359, 2013.

[14] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[15] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787 – 800, 2007.

[16] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, and R. Cowie, "Data-driven clustering in emotional space for affect recognition using discriminatively trained lstm networks," in *INTERSPEECH*, 2009.

[17] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768 – 785, 2011.

[18] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?" in *INTERSPEECH-2011*, 2011.

[19] B. Schuller, "Recognizing affect from linguistic information in 3d continuous space," *Affective Computing, IEEE Transactions on*, vol. 2, no. 4, pp. 192 –205, 2011.

[20] P. Fewzee and F. Karray, "Emotional speech: A spectral analysis," in *INTERSPEECH*, 2012, pp. 2238–2241.

[21] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *14th Int. Joint Conf. on AI*, vol. 2, 1995, pp. 1137–1143.

[22] J. Rodriguez, A. Perez, and J. Lozano, "Sensitivity analysis of k-fold cross validation in prediction error estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 569–575, 2010.

[23] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. Affect. Comp.*, vol. 4, no. 1, pp. 15–33, 2013.