Electronic Journal of Statistics
Vol. 12 (2018) 960–984
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1403

Supervised dimensionality reduction via distance correlation maximization

Praneeth Vepakomma*

Department of Statistics
Rutgers University, New Brunswick, U.S.A
and Motorola Solutions, Atlanta, USA
e-mail: praneeth@scarletmail.rutgers.edu

Chetan Tonde* and Ahmed Elgammal

Department of Computer Science
Rutgers University, New Brunswick, U.S.A
e-mail: cjtonde@scarletmail.rutgers.edu; elgammal@cs.rutgers.edu

Abstract: In our work, we propose a novel formulation for supervised dimensionality reduction based on a nonlinear dependency criterion called Statistical Distance Correlation, (Székely et al., 2007). We propose an objective which is free of distributional assumptions on regression variables and regression model assumptions. Our proposed formulation is based on learning a low-dimensional feature representation $\mathbf{z}$, which maximizes the squared sum of Distance Correlations between low-dimensional features $\mathbf{z}$ and response y, and also between features $\mathbf{z}$ and covariates $\mathbf{x}$. We propose a novel algorithm to optimize our proposed objective using the Generalized Minimization Maximization method of (Parizi et al., 2015). We show superior empirical results on multiple datasets proving the effectiveness of our proposed approach over several relevant state-of-the-art supervised dimensionality reduction methods.

Keywords and phrases: Distance correlation, multivariate statistical independence, minorization maximization, supervised dimensionality reduction, fixed point iteration, optimization, representation learning.

Received November 2016.

1. Introduction

Rapid developments of imaging technology, microarray data analysis, computer vision, neuroimaging, hyperspectral data analysis and many other applications call for the analysis of high-dimensional data. The problem of supervised dimensionality reduction is concerned with finding a low-dimensional representation of data such that this representation can be effectively used in a supervised learning task. Such representations help in providing a meaningful interpretation and visualization of the data, and also help to prevent overfitting when the number of dimensions greatly exceeds the number of samples, thus working as a form of regularization. In this paper we focus on supervised dimensionality

*Equal contribution.

reduction in the regression setting where we consider the problem of predicting a univariate response $y_i \in \mathbb{R}$ from a vector of continuous covariates $\mathbf{x}_i \in \mathbb{R}^p$, for $i = 1$ to n.

Sliced Inverse Regression (SIR) of Li (1991); Lue (2009); Szretter and Yohai (2009) is one of the earliest developed supervised dimensionality reduction techniques and is a seminal work that introduced the concept of a central subspace that we now describe. This technique aims to find a subspace given by the column space of a $p \times d$ matrix $\mathbf{B}$ with $d << p$ such that $\mathbf{y} \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^T\mathbf{X}$ where $\perp\!\!\!\perp$ indicates statistical independence. Under mild conditions the intersection of all such dimension reducing subspaces is itself a dimension reducing subspace, and is called the central subspace (Cook, 1996). SIR aims to estimate this central subspace. Sliced Average Variance Estimation (SAVE) of Shao et al. (2009) and Shao et al. (2007) is another early method that can be used to estimate the central subspace. SIR uses a sample version of the first conditional moment $\mathbf{E}\mathbf{X} \mid Y$ to construct an estimator of this subspace and SAVE uses the sample first and second conditional moments to estimate it. Likelihood Acquired Directions (LAD) of Cook and Forzani (2009) is a technique that obtains the maximum likelihood estimator of the central subspace under assumptions of conditional normality of the predictors given the response. Like LAD, methods SIR and SAVE rely on elliptical distributional assumptions like Gaussianity of the data.

More recently developed methods do not require any distributional assumptions on the marginal distribution of $\mathbf{x}$ or on the conditional distribution of y. The authors of Gradient Based Kernel Dimension Reduction (gKDR), Fukumizu and Leng (2014), use an equivalent formulation of the conditional independence relation $\mathbf{y} \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^T\mathbf{X}$ using conditional cross-covariance operators and aim to find a $\mathbf{B}$ that maximizes the mutual information $I(\mathbf{B^T X}, \mathbf{y})$. In this work, the authors use Gaussian kernels to provide equivalent characterizations of conditional independence using sample estimators of cross-covariance operators.

Sufficient Component Analysis (SCA) of Yamada et al. (2011) is another technique where the $\mathbf{B}$ is also learnt using a dependence criterion. SCA aims to maximize the least-squares mutual information given by $SMI(Z, Y) = \frac{1}{2}\int\int(\frac{p_{zy}(z,y)}{p_z(z)p_y(y)} - 1)^2 dzdy$ between the projected features $\mathbf{Z} = \mathbf{B}^T\mathbf{X}$ and the response. This is done under orthonormal constraints over $\mathbf{B}$, and the optimal solution is found by approximating $\frac{p_{zy}(z,y)}{p_z(z)p_y(y)}$ using method of density ratio estimation (Sugiyama et al., 2012; Vapnik et al., 2015), and also an analytical closed form solution for the minima is obtained. In Suzuki and Sugiyama (2013) (LSDR), the authors optimize this objective using a natural gradient based iterative solution on the Steifel manifold $\mathbb{S}_d^m(\mathbb{R})$ via a line search along the geodesic in the direction of the natural gradient (Amari, 1998; Nishimori and Akaho, 2005). In our work, we show benefits of Distance Correlation as a criterion for supervised low-dimensional feature learning.

Our contribution in this paper is as follows: We propose a new formulation for supervised dimensionality reduction that is based on a dependency criterion called Distance Correlation, (Szekely et al., 2007). This setup is free of distribu-

tional, as well as regression model assumptions. The novelty in our formulation is that we do not restrict the transformation from $\mathbf{x}$ to $\mathbf{z}$ to be linear, as in case many of the above techniques.

To further add to this, the recent work of Sheng and Yin (2016) looks at sufficient dimensionality reduction through linear projections using distance covariance and Xin Chen and Zou (2015) looks at goodness-of-fit tests with distance covariance in the context of sufficient dimensionality reduction. In addition, Li et al. (2012); Kong et al. (2015); Berrendero José R and Torrecilla (2014) have used Distance Correlation as a criterion for feature selection in a regression setting.

In our work we use the following notation: The spectral radius of a matrix $\mathbf{M}$ is denoted by $\lambda_{max}(\mathbf{M})$, i^{th} eigenvalue by $\lambda_i(\mathbf{M})$, and i^{th} generalized eigenvalue $\mathbf{Ax} = \lambda_i\mathbf{Bx}$ by $\lambda_i(\mathbf{A}, \mathbf{B})$. Moreover, $\lambda_{max}(M)$ $(\lambda_{max}(\mathbf{A}, \mathbf{B}))$, and $\lambda_{max}(M)$ $(\lambda_{min}(\mathbf{A}, \mathbf{B}))$ respectively, the maximum and minimum eigenvalues (generalized eigenvalues) of matrices $\mathbf{M}$, $\mathbf{A}$ and $\mathbf{B}$. We use the usual partial ordering for symmetric matrices: $\mathbf{A} \succeq \mathbf{B}$ means $\mathbf{A} - \mathbf{B}$ is positive semidefinite; similarly for the relationships $\succeq, \prec, \succ$. The norm $\|\cdot\|$ will be either the Euclidean norm for vectors or the norm that it induces for matrices, unless otherwise specified.

2. Distance correlation

Distance Correlation introduced by Szekely et al. (2007) and Székely et al. (2009); Székely and Rizzo (2012, 2013) is a measure of nonlinear dependencies between random vectors of arbitrary dimensions. We describe below α-distance covariance which is an extended version of standard distance covariance for $\alpha = 1$.

Definition 2.1. Distance Covariance (Székely et al., 2007), α-dCov: Distance covariance between random variables $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^m$ with finite first moments is a nonnegative number given by

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

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

$$w(t, s) = (C(p, \alpha)C(q, \alpha)|t|_p^{\alpha+p}|s|_q^{\alpha+q})^{-1}$$

with $C(d, \alpha) = \frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma((\alpha+d)/2)}$.

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

Definition 2.2. Distance Correlation (Székely et al., 2007) (α-dCorr): 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 defined as

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

The Distance Correlation defined above has the following interesting properties; 1) $\rho^2(\mathbf{x},\mathbf{x})$ is defined for arbitrary dimensions of $\mathbf{x}$ and $\mathbf{y}$, 2) $\rho^2(\mathbf{x},\mathbf{y}) = 0$ if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent, and 3) $\rho^2(\mathbf{x},\mathbf{y})$ satisfies the relation $0 \leq \rho^2(\mathbf{x},\mathbf{y}) \leq 1$. In our work, we use α-Distance Covariance with $\alpha = 2$ and in the following paper for simplicity just refer to it as Distance Correlation.

We define sample version of distance covariance given i.i.d. samples $\{(\mathbf{x}_k,\mathbf{y}_k)|\ k = 1,2,\ldots,n\}$ sampled 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 $k,l \in \{1,2,\ldots,n\}$. We then make their row and column sums zero to obtain $\widehat{\mathbf{E}}_\mathbf{X}$ and $\widehat{\mathbf{E}}_\mathbf{Y}$ respectively by multiplying with a centering matrix Borg and Groenen (2005) $\mathbf{J}$ given by $\mathbf{J} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ on both sides as $\widehat{\mathbf{E}}_\mathbf{X} = \mathbf{J}\mathbf{E}_\mathbf{X}\mathbf{J}$ and $\widehat{\mathbf{E}}_\mathbf{Y} = \mathbf{J}\mathbf{E}_\mathbf{Y}\mathbf{J}$ Now the sample distance correlation (for $\alpha = 2$) is hence defined as follows:

Definition 2.3. Sample Distance Correlation (Székely et al., 2007): 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 $\widehat{\mathbf{E}}_\mathbf{Y}$, the squared sample distance correlation is defined as,

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

and equivalently sample distance correlation is given by

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

3. Laplacian formulation of sample distance correlation

In this section, we propose a Laplacian formulation of sample distance covariance and sample distance correlation which we later use to propose our objective function used for supervised dimensionality reduction (SDR).

A graph Laplacian version of sample distance correlation can be obtained as follows:

Lemma 3.1. *Given matrices of squared Euclidean distances* $\mathbf{E_X}$ *and* $\mathbf{E_Y}$ *and Laplacians* $\mathbf{L_X}$ *and* $\mathbf{L_Y}$ *formed over adjacency matrics* $\widehat{\mathbf{E}}_\mathbf{X}$ *and* $\widehat{\mathbf{E}}_\mathbf{Y}$*, the square of sample distance correlation* $\hat{\rho}^2(\mathbf{X},\mathbf{Y})$ *is given by*

$$\hat{\rho}^2(\mathbf{X},\mathbf{Y}) = \frac{\mathbf{Tr}\left(\mathbf{X}^T\mathbf{L_Y}\mathbf{X}\right)}{\sqrt{\mathbf{Tr}\left(\mathbf{Y}^T\mathbf{L_Y}\mathbf{Y}\right)\mathbf{Tr}\left(\mathbf{X}^T\mathbf{L_X}\mathbf{X}\right)}}. \tag{3.1}$$

Proof. Given matrices $\widehat{\mathbf{E}}_{\mathbf{X}}$, $\widehat{\mathbf{E}}_{\mathbf{Y}}$, and column centered matrices $\widetilde{\mathbf{X}}$, $\widetilde{\mathbf{Y}}$, from result of Torgerson (1952) we have that $\widehat{\mathbf{E}}_{\mathbf{X}} = -2\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$ and $\widehat{\mathbf{E}}_{\mathbf{Y}} = -2\widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T$. In the problem of multidimensional scaling (MDS) (Borg and Groenen, 2005), we know for a given adjacency matrix say $\mathbf{W}$ and a Laplacian matrix $\mathbf{L}$,

$$\mathbf{Tr}\left(\mathbf{X}^T\mathbf{L}\mathbf{X}\right) = \frac{1}{2}\sum_{i,j}[\mathbf{W}]_{ij}[\mathbf{E}_{\mathbf{X}}]_{i,j}. \tag{3.2}$$

Now for the Laplacian $\mathbf{L} = \mathbf{L}_{\mathbf{X}}$ and adjacency matrix $\mathbf{W} = \widehat{\mathbf{E}}_{\mathbf{Y}}$ we can represent $\mathbf{Tr}\left(\mathbf{X}^T\mathbf{L}_{\mathbf{Y}}\mathbf{X}\right)$ in terms of $\widehat{\mathbf{E}}_{\mathbf{Y}}$ as follows,

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

From the fact $[\mathbf{E}_{\mathbf{X}}]_{i,j} = (\langle\widetilde{\mathbf{x}}_i,\widetilde{\mathbf{x}}_i\rangle + \langle\widetilde{\mathbf{x}}_j,\widetilde{\mathbf{x}}_j\rangle - 2\langle\widetilde{\mathbf{x}}_i,\widetilde{\mathbf{x}}_j\rangle)$, and also $\widehat{\mathbf{E}}_{\mathbf{X}} = -2\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T$ we get

$$\begin{aligned}\mathbf{Tr}\left(\mathbf{X}^T\mathbf{L}_{\mathbf{Y}}\mathbf{X}\right) &= -\frac{1}{4}\sum_{i,j=1}^{n}[\widehat{\mathbf{E}}_{\mathbf{Y}}]_{i,j}([\widehat{\mathbf{E}}_{\mathbf{X}}]_{i,i} + [\widehat{\mathbf{E}}_{\mathbf{X}}]_{j,j} - 2[\widehat{\mathbf{E}}_{\mathbf{X}}]_{i,j}) \\ &= \frac{1}{2}\sum_{i,j}[\widehat{\mathbf{E}}_{\mathbf{X}}]_{i,j}[\widehat{\mathbf{E}}_{\mathbf{Y}}]_{i,j} - \frac{1}{4}\sum_{j}^{n}[\widehat{\mathbf{E}}_{\mathbf{X}}]_{j,j}\sum_{i}^{n}[\widehat{\mathbf{E}}_{\mathbf{Y}}]_{i,j} \\ &\quad - \frac{1}{4}\sum_{i}^{n}[\widehat{\mathbf{E}}_{X}]_{i,i}\sum_{j}^{n}[\widehat{\mathbf{E}}_{\mathbf{Y}}]_{i,j}\end{aligned}$$

Since $\widehat{\mathbf{E}}_{\mathbf{X}}$ and $\widehat{\mathbf{E}}_{\mathbf{Y}}$ are double centered matrices $\sum_{i=1}^{n}[\widehat{\mathbf{E}}_{\mathbf{Y}}]_{i,j} = \sum_{j=1}^{n}[\widehat{\mathbf{E}}_{\mathbf{Y}}]_{i,j} = 0$ it follows that

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

It also follows that

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

Similarly, we can express the sample distance covariance using Laplacians $\mathbf{L}_{\mathbf{X}}$ and $\mathbf{L}_{\mathbf{Y}}$ as

$$\hat{\nu^2}(\mathbf{X},\mathbf{Y}) = \left(\frac{2}{n^2}\right)\mathbf{Tr}\left(\mathbf{X}^T\mathbf{L}_{\mathbf{Y}}\mathbf{X}\right) = \left(\frac{2}{n^2}\right)\mathbf{Tr}\left(\mathbf{Y}^T\mathbf{L}_{\mathbf{X}}\mathbf{Y}\right).$$

The sample distance variances can be expressed as $\hat{\nu}^2(\mathbf{X},\mathbf{X}) = \left(\frac{2}{n^2}\right)\mathbf{Tr}\left(\mathbf{X}^T\mathbf{L}_{\mathbf{X}}\mathbf{X}\right)$ and $\hat{\nu}^2(\mathbf{Y},\mathbf{Y}) = \left(\frac{2}{n^2}\right)\mathbf{Tr}\left(\mathbf{Y}^T\mathbf{L}_{\mathbf{Y}}\mathbf{Y}\right)$ substituting back into expression of sample distance correlation above we get Equation 3.1. □

4. Framework

4.1. Problem statement

The goal in supervised dimensionality reduction (SDR) is to learn a low dimensional representation $\mathbf{Z} \in \mathbb{R}^{n \times p}$ of input features $\mathbf{X} \in \mathbb{R}^{n \times d}$ so as to predict the response $\mathbf{y} \in \mathbb{R}^{n \times 1}$ from $\mathbf{Z}$.

In our proposed formulation, we use aforementioned Laplacian based sample distance correlation to measure dependencies between variables. We propose to maximize dependencies between the low dimensional features $\mathbf{Z}$ and response vector $\mathbf{y}$, and also low dimensional features $\mathbf{Z}$ with input features $\mathbf{X}$. Our objective is to maximize the sum of squares of these two sample distance correlations which is given by:

$$f(\mathbf{Z}) = \hat{\rho}^2(\mathbf{X}, \mathbf{Z}) + \hat{\rho}^2(\mathbf{Z}, \mathbf{y}) \tag{4.1}$$

$$f(\mathbf{Z}) = \frac{\mathbf{Tr}\left(\mathbf{Z}^T \mathbf{L}_\mathbf{X} \mathbf{Z}\right)}{\sqrt{\mathbf{Tr}\left(\mathbf{X}^T \mathbf{L}_\mathbf{X} \mathbf{X}\right) \mathbf{Tr}\left(\mathbf{Z}^T \mathbf{L}_\mathbf{Z} \mathbf{Z}\right)}} + \frac{\mathbf{Tr}\left(\mathbf{Z}^T \mathbf{L}_\mathbf{y} \mathbf{Z}\right)}{\sqrt{\mathbf{Tr}\left(\mathbf{y}^T \mathbf{L}_\mathbf{y} \mathbf{y}\right) \mathbf{Tr}\left(\mathbf{Z}^T \mathbf{L}_\mathbf{Z} \mathbf{Z}\right)}}. \tag{4.2}$$

On simplification we get the following optimization problem which we refer to as **Problem (P)**.

$$\max_{\mathbf{Z}} \qquad f(\mathbf{Z}) = \frac{\mathbf{Tr}\left(\mathbf{Z}^T \mathbf{S}_{\mathbf{X},\mathbf{y}} \mathbf{Z}\right)}{\sqrt{\mathbf{Tr}\left(\mathbf{Z}^T \mathbf{L}_\mathbf{Z} \mathbf{Z}\right)}} \qquad \textbf{Problem (P)}$$

where $k_X = \frac{1}{\sqrt{\mathbf{Tr}(\mathbf{X}^T \mathbf{L}_\mathbf{X} \mathbf{X})}}$, $k_Y = \frac{1}{\sqrt{\mathbf{Tr}(\mathbf{y}^T \mathbf{L}_\mathbf{y} \mathbf{y})}}$ are constants, and $\mathbf{S}_{\mathbf{X},\mathbf{y}} = k_X \mathbf{L}_\mathbf{X} + k_Y \mathbf{L}_\mathbf{y}$.

4.1.1. Motivation

To elucidate on the intuition behind the two additive terms $\hat{\rho}^2(\mathbf{X}, \mathbf{Z})$ and $\hat{\rho}^2(\mathbf{Z}, \mathbf{y})$ in above objective $\hat{\rho}^2(\mathbf{X}, \mathbf{Z}) + \hat{\rho}^2(\mathbf{Z}, \mathbf{y})$, we first point that for a special case of learning a $\mathbf{Z}_{\texttt{Train}}$ such that $\mathbf{Z}_{\texttt{Train}} = \mathbf{y}_{\texttt{Train}}$, given a training data of $\mathbf{X}_{\texttt{Train}}, \mathbf{y}_{\texttt{Train}}$, we see that although the term $\hat{\rho}^2(\mathbf{Z}_{\texttt{Train}}, \mathbf{y}_{\texttt{Train}})$ is maximum it happens to be the case that $\mathbf{Z}_{\texttt{Train}}$ is a grossly over-fitted representation of the training data. That is, it does not aid in learning a reasonable $\mathbf{Z}_{\texttt{Test}}$ to predict an out-of-sample response $\mathbf{Y}_{\texttt{Test}}$.

Optimizing our proposed objective function facilitates the learning of a reasonable out-of-sample $\mathbf{Z}_{\texttt{Test}}$ through two explicit supervised machine learning maps; one that maps from $\mathbf{X}_{\texttt{Train}}$ to $\mathbf{Z}_{\texttt{Train}}$ to help predict $\mathbf{Z}_{\texttt{Test}}$ and another that maps from $\mathbf{Z}_{\texttt{Train}}$ to $\mathbf{Y}_{\texttt{Train}}$ help predict $\mathbf{Y}_{\texttt{Test}}$ from the predicted $\mathbf{Z}_{\texttt{Test}}$. Popular supervised machine learning techniques could be used to learn the two maps above.

We also show through experiments later in this paper that our proposed iterative solution of $\mathbf{Z}$ for this objective tries to increase $\hat{\rho}^2(\mathbf{Z}, \mathbf{Y})$ at a greater rate while $\hat{\rho}^2(\mathbf{Z}, \mathbf{X})$ reduces at a relatively slower rate with respect to iterations. We also visualize the plots of variables in $\mathbf{Z}$ vs. corresponding variables in $\mathbf{X}$

and the variables in $\mathbf{Z}$ vs. $\mathbf{Y}$ later in this paper at different iterations to further show this effect.

Another parallel, yet motivating line of work on this topic is the popular "Information Bottleneck Method" introduced and studied in (Tishby Naftali and William, 1999; Chechik Gal and Yair, 2005; Noam, 2002) which tries to find a compressed representation of $\mathbf{X}$ while also preserving relevant information about $\mathbf{Y}$. This was used very recently to further explain the theoretical underpinnings of deep learning in Ravid and Naftali (2017).

4.2. Algorithm

In the proposed problem (**Problem (P)**), we observe that numerator of our objective is convex while denominator is non-convex due to the presence of a square root and a Laplacian term $\mathbf{L_Z}$ nonlinearly dependent on $\mathbf{Z}$. Hence, this makes direct optimization of this objective practically infeasible. So to optimize **Problem (P)**, we present a surrogate objective **Problem (Q)** which lower bounds our proposed original objective. We maximize this lower bound with respect to $\mathbf{Z}$ and show that optimizing this surrogate objective **Problem (Q)** (lower bound), also maximizes the proposed objective in **Problem (P)**. We do so by utlizing the Generalized Minorization-Maximization (G-MM) framework of Parizi et al. (2015).

The G-MM framework of Parizi et al. (2015) is an extension of the well known MM framework of Lange et al. (2000). It removes the equality constraint between both objectives at every iteration $\mathbf{Z}_k$, except at initialization step $\mathbf{Z}_0$. This allows the use of a broader class of surrogate objective functions.

The surrogate lower bound objective is as follows,

$$\max_{\mathbf{Z}} \qquad g(\mathbf{Z}, \mathbf{M}) = \frac{\mathbf{Tr}\left(\mathbf{Z}^T \mathbf{S}_{\mathbf{X},\mathbf{y}} \mathbf{Z}\right)}{\mathbf{Tr}\left(\mathbf{Z}^T \mathbf{L}_{\mathbf{M}} \mathbf{Z}\right)} \qquad \textbf{Problem (Q)}$$

where $\mathbf{M} \in \mathbb{R}^{n \times d}$ belongs to the set of column-centered matrices.

The surrogate problem (**Problem (Q)**) is convex in both its numerator and denominator for a fixed auxiliary variable $\mathbf{M}$. Theorem 4.1 provides the required justification that under certain conditions, maximizing the surrogate **Problem (Q)** also maximizes the proposed objective **Problem (P)** .

An outline of the strategy for optimization is as follows:

a) **Initialize**: Initialize $\mathbf{Z}_0 = \left[c\mathbf{J}_d, \mathbf{0}^T_{(n-d)\times d}\right]^T$, a column-centered matrix where $c = \frac{1}{\sqrt[4]{2(d-1)}}$ and $\mathbf{J}_d \in \mathbb{R}^{d \times d}$ is a centering matrix. This is motivated by statement 1) in proof of Theorem 4.1.
b) **Optimize**: Maximize the surrogate lower bound $\mathbf{Z}_{k+1} = \arg\max g(\mathbf{Z}, \mathbf{Z}_k)$ (See section 5).
c) **Rescaling**: Rescale $\mathbf{Z}_{k+1} \leftarrow \kappa \mathbf{Z}_{k+1}$ such that $\mathbf{Tr}\left(\mathbf{Z}_{k+1} \mathbf{L}_{\mathbf{Z}_{k+1}} \mathbf{Z}_{k+1}\right)$ is greater than one. This is motivated by proof of statement 3) of Theorem 4.1, and also the fact that $g(\mathbf{Z}, \mathbf{M}) = g(\kappa \mathbf{Z}, \mathbf{M})$ and $f(\mathbf{Z}) = f(\kappa \mathbf{Z})$ for any scalar κ.
d) Repeat step b and c above until convergence.

Theorem 4.1. *Under above strategy, maximizing the surrogate* ***Problem Q*** *also maximizes* ***Problem P****.*

Proof. For convergence it is enough for us to show the following, (Parizi et al., 2015):

1. $f(\mathbf{Z}_0) = g(\mathbf{Z}_0, \mathbf{Z}_0)$ for $\mathbf{Z}_0 = \left[c\mathbf{J}_d, \mathbf{0}^T_{(n-d)\times d}\right]^T$ and $c = \frac{1}{\sqrt[4]{2(d-1)}}$,
2. $g(\mathbf{Z}_{k+1}, \mathbf{Z}_k) \geq g(\mathbf{Z}_k, \mathbf{Z}_k)$ and,
3. $f(\mathbf{Z}_{k+1}) \geq g(\mathbf{Z}_{k+1}, \mathbf{Z}_k)$

To prove statement 1, for $\mathbf{Z}_0 = \left[c\mathbf{J}_d, \mathbf{0}^T_{(n-d)\times d}\right]^T$, we observe that $\mathbf{Z}_0$ column-centered, $\mathbf{L}_{\mathbf{Z}_0} = 2\mathbf{Z}_0\mathbf{Z}_0^T$ and $\mathbf{Z}_0^T\mathbf{Z}_0 = c^2\mathbf{J}_d$. Hence we get $\mathbf{Tr}\left(\mathbf{Z}_0^T\mathbf{Z}_0\mathbf{L}_{\mathbf{Z}_0}\mathbf{Z}_0\right) = c^4\mathbf{Tr}\left(2\mathbf{J}_d\right) = c^4 2(d-1) = 1$. This proves the required statement $f(\mathbf{Z}_0) = g(\mathbf{Z}_0, \mathbf{Z}_0) = \mathbf{Tr}\left(\mathbf{Z}_0^T\mathbf{L}_{\mathbf{Z}_0}\mathbf{Z}_0^T\right)$.

Statement 2 follows from the optimization $\mathbf{Z}_{k+1} = \arg\max g(\mathbf{Z}, \mathbf{Z}_k)$. To prove statement 3 we have to show that

$$\frac{\mathbf{Tr}\left(\mathbf{Z}_{k+1}^T\mathbf{S}_{\mathbf{X},\mathbf{y}}\mathbf{Z}_{k+1}\right)}{\sqrt{\mathbf{Tr}\left(\mathbf{Z}_{k+1}^T\mathbf{L}_{\mathbf{Z}_{k+1}}\mathbf{Z}_{k+1}\right)}} \geq \frac{\mathbf{Tr}\left(\mathbf{Z}_{k+1}^T\mathbf{S}_{\mathbf{X},\mathbf{y}}\mathbf{Z}_{k+1}\right)}{\mathbf{Tr}\left(\mathbf{Z}_{k+1}^T\mathbf{L}_{\mathbf{Z}_k}\mathbf{Z}_{k+1}\right)}.$$

Since numerators on both sides are equal, it is enough for us to show that

$$\sqrt{\mathbf{Tr}\left(\mathbf{Z}_{k+1}^T\mathbf{L}_{\mathbf{Z}_{k+1}}\mathbf{Z}_{k+1}\right)} \leq \mathbf{Tr}\left(\mathbf{Z}_{k+1}^T\mathbf{L}_{\mathbf{Z}_k}\mathbf{Z}_{k+1}\right).$$

Now from Lemma A.4 we have $\mathbf{Tr}\left(\mathbf{Z}_{k+1}^T\mathbf{L}_{\mathbf{Z}_{k+1}}\mathbf{Z}_{k+1}\right) \leq \mathbf{Tr}\left(\mathbf{Z}_{k+1}^T\mathbf{L}_{\mathbf{Z}_k}\mathbf{Z}_{k+1}\right)$. It follows from the rescaling step (step c) of the optimization strategy that the left hand side $\mathbf{Tr}\left(\mathbf{Z}_{t+1}\mathbf{L}_{\mathbf{Z}_{t+1}}\mathbf{Z}_{t+1}\right)$ is always greater that one, and so taking square root of it implies $\sqrt{\mathbf{Tr}\left(\mathbf{Z}_{t+1}\mathbf{L}_{\mathbf{Z}_{t+1}}\mathbf{Z}_{t+1}\right)} \leq \mathbf{Tr}\left(\mathbf{Z}_{t+1}^T\mathbf{L}_{\mathbf{Z}_t}\mathbf{Z}_{t+1}\right)$. □

We summarize all of the above steps in Algorithm 4.1 below and section 5 further describes optimization algorithm to solve **Problem (Q)** required by it.

Algorithm 4.1 DISCOMAX

Require: Initialize $\mathbf{Z}_0 = \left[c\mathbf{J}_d, \mathbf{0}^T_{(n-d)\times d}\right]^T$, a column-centered matrix where $c = \frac{1}{\sqrt[4]{2(d-1)}}$, $k \leftarrow 0$

Ensure: $\mathbf{Z}^* = \arg\max_{\mathbf{Z}} f(\mathbf{Z})$

1: **repeat**
2: Solve,

$$\mathbf{Z}_{k+1} = \arg\max_{\mathbf{Z}} g(\mathbf{Z}, \mathbf{Z}_k) \qquad \textbf{Problem (Q)}$$

3: Rescale $\mathbf{Z}_{k+1} \leftarrow \kappa\mathbf{Z}_{k+1}$ such that $\mathbf{Tr}\left(\mathbf{Z}_{k+1}^T\mathbf{L}_{\mathbf{Z}_{k+1}}\mathbf{Z}_{k+1}\right) \geq 1$
4: $k = k + 1$
5: **until** $\|\mathbf{Z}_{k+1} - \mathbf{Z}_k\|^2 < \epsilon$
6: $\mathbf{Z}^* = \mathbf{Z}_{k+1}$
7: **return** $\mathbf{Z}^*$

5. Optimization

In this section, we propose a framework for optimizing the surrogate objective $g(\mathbf{Z}, \mathbf{M})$, referred to as **Problem (Q)**, for a fixed $\mathbf{M} = \mathbf{Z}_k$. We observe that for a given value of $\mathbf{M}$, $g(\mathbf{Z}, \mathbf{M})$ is a ratio of two convex functions. To solve this, we convert this maximization problem to an equivalent minimization problem $h(\mathbf{Z}, \mathbf{M})$, by taking its reciprocal (Schaible, 1976). This allows us to utilize the Quadratic Fractional Programming Problem (QFPP) framework of Dinkelbach (1967) and Zhang (2008) to minimize $h(\mathbf{Z}, \mathbf{M})$. We refer to this new minimization problem as **Problem (R)**. It is stated below.

$$\min_{\mathbf{Z}} \qquad h(\mathbf{Z}, \mathbf{M}) = \frac{\mathbf{Tr}\left(\mathbf{Z}^T \mathbf{L}_{\mathbf{M}} \mathbf{Z}\right)}{\mathbf{Tr}\left(\mathbf{Z}^T \mathbf{S}_{\mathbf{X},\mathbf{y}} \mathbf{Z}\right)} \qquad \textbf{Problem (R)} \tag{5.1}$$

where $\mathbf{M} = \mathbf{Z}_k$.

In his seminal work Dinkelbach (1967) and later Zhang (2008) proposed a novel framework to solve constrained QFP problems by converting it to an equivalent parametric optimization problem, by introducing a scalar parameter $\alpha \in \mathbb{R}$. We utilize this equivalence proposed to defined new parametric problem, **Problem (S)**. The solution involves a search over the scalar parameter α while repeatedly solving **Problem (S)** to get the required solution $\mathbf{Z}_{k+1}$. This search process continues until values of α converge.

In a nutshell, Dinkelbach (1967) and Zhang (2008) frameworks suggest the following optimizations are equivalent:

Problem (R)
$\underset{\mathbf{z} \in \mathbb{R}^d}{\text{minimize}}\ h(\mathbf{z}) = \frac{f_1(\mathbf{z})}{f_2(\mathbf{z})}$

$\Longleftrightarrow$

Problem (S)
$\underset{\mathbf{z} \in \mathbb{R}^d}{\text{minimize}}\ H(\mathbf{z}; \alpha^*) = f_1(\mathbf{z}) - \alpha^* f_2(\mathbf{z})$
for some $\alpha^* \in \mathbb{R}$

where $f_i(\mathbf{z}) := \mathbf{z}_i^T \mathbf{A}_i \mathbf{z} - 2\mathbf{b}_i \mathbf{z} + c_i$ with $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{n \times n}$, $\mathbf{b}_1, \mathbf{b}_2 \in \mathbf{R}^n$, and $c_1, c_2 \in \mathbb{R}$. $\mathbf{A}_1$ and $\mathbf{A}_2$ are symmetric with $f_2(\mathbf{x}) > 0$ over some $\mathbf{z} \in \mathcal{Z}$.

To see the equivalence of $h(\mathbf{Z}, \mathbf{M})$ in **Problem (R)** to $h(\mathbf{z})$ above we observe that: $\mathbf{A}_1 = \mathbf{I}_n \otimes \mathbf{L}_{\mathbf{M}}$, $\mathbf{A}_2 = \mathbf{I}_n \otimes \mathbf{S}_{\mathbf{X},\mathbf{y}}$, $c_i = c_2 = 0$, and $\mathbf{b}_1 = \mathbf{b}_2 = \mathbf{0}$. Also, due to positive definiteness of $\mathbf{A}_i$, $f_i(\mathbf{z})$ is positive[1], and $f(\mathbf{z}_i) > 0$. Using this setup for $h(\mathbf{Z}, \mathbf{M})$ we get,[2]

$$\min_{\mathbf{Z}} \qquad h(\mathbf{Z}, \mathbf{M}) = \frac{\mathbf{vec}\,(\mathbf{Z})^T\,(\mathbf{I}_n \otimes \mathbf{L}_{\mathbf{M}})\mathbf{vec}\,(\mathbf{Z})}{\mathbf{vec}\,(\mathbf{Z})^T\,(\mathbf{I}_n \otimes \mathbf{S}_{\mathbf{X},\mathbf{y}})\mathbf{vec}\,(\mathbf{Z})} \tag{5.2}$$

In subsection 5.1 we propose a Golden Section Search (Kiefer, 1953) based algorithm (Algorithm 5.1) which utilizes concavity property of $H(\mathbf{Z}; \alpha)$ with respect to α to locate the best α^*. During this search we repeatedly solve **Problem (S)** starting with an intial interval $0 = \alpha_l \leq \alpha \leq \alpha_u = \lambda_{min}(\mathbf{L}_{\mathbf{M}}, \mathbf{S}_{\mathbf{X},\mathbf{y}})$ for a

[1] In case of $\mathbf{A}_i$ is semi-definite we regularize by adding $\mathbf{A}_i + \epsilon\mathbf{I}$ so that $\mathbf{A}_i \succ 0$

[2] $\otimes$ indicates kronecker product. $\mathbf{vec}\,(\mathbf{Z})$ denotes column vectorization of matrix $\mathbf{Z}$.

fixed $\mathbf{M}$, then at each step shorten the search interval by moving upper and lower limits closer to each other. We continue until convergence to α^*. The choice of the upper limit of $\alpha_u = \lambda_{min}(\mathbf{L_M}, \mathbf{S_{X,y}})$ is motivated by proof of Lemma A.2.

To solve **Problem (S)** for a given α, we propose an iterative algorithm in subsection 5.2 (Algorithm 5.2). It uses the classical Majorization-Minimization framework of Lange (2013).

5.1. Golden Section Search

Dinkelbach (1967) and Zhang (2008) showed the following properties of the objective[3] $H(\alpha)$ with respect to α, for a fixed $\mathbf{Z}$.

Theorem 5.1. *Let* $G\colon \mathbb{R} \to \mathbb{R}$ *be defined as*

$$G(\alpha) = \min_{\mathbf{Z}} H(\mathbf{Z}; \alpha) = \min_{\mathbf{Z}} \left\{ \mathbf{Tr}\left(\mathbf{Z}^T \mathbf{L_M} \mathbf{Z}\right) - \alpha \mathbf{Tr}\left(\mathbf{Z}^T \mathbf{S_{X,y}} \mathbf{Z}\right) \right\}$$

as derived from ***Problem (S)****, then following statements hold true.*

1. *G is continuous at any $\alpha \in \mathbb{R}$.*
2. *G is concave over $\alpha \in \mathbb{R}$.*
3. *$G(\alpha) = 0$, has a unique solution α^*.*

Algorithm 5.1 exploits the concavity property of $G(\alpha)$ to perform a Golden Section Search over α. Subsection 5.2 provides an iterative Majorization-Minimization algorithm (Algorithm 5.2) to solve this minimization problem **Problem (S)**.

5.2. Distance correlation maximization algorithm

Algorithm 5.2 gives a iterative fixed point algorithm which solves **Problem (S)**. Theorem 5.2 provides a fixed point iterate used to minimize $H(\mathbf{Z}, \alpha)$ with respect to $\mathbf{Z}$ for a given α. The fixed point iterate[4] $\mathbf{Z}_{t+1} = \mathbf{H}\mathbf{Z}_t$ minimizes **Problem (S)** and a monotonic convergence is assured by the Majorization-Minimization result of Lange (2013). Theorem 5.2 below derives the fixed point iterate used in Algorithm 5.2.

Theorem 5.2. *For a fixed γ^2 (Lemma A.1), some α (Lemma A.2) and*

$$\mathbf{H} = \left(\gamma^2 \mathbf{D}_X - \alpha \mathbf{S_{X,y}}\right)^{\dagger} \left(\gamma^2 \mathbf{D}_X - \mathbf{L_M}\right)$$

the iterate $\mathbf{Z}_t = \mathbf{H}\mathbf{Z}_{t-1}$ monotonically minimizes the objective,

$$F(\mathbf{Z}; \alpha) = \mathbf{Tr}\left(\mathbf{Z}^T \mathbf{L_M} \mathbf{Z}\right) - \alpha \mathbf{Tr}\left(\mathbf{Z}^T \mathbf{S_{X,y}} \mathbf{Z}\right) \tag{5.3}$$

[3] For a fixed $\mathbf{Z}$ and variable argument α we denote $H(\mathbf{Z}; \alpha)$ as $H(\alpha)$.
[4] We use the subscript t to indicate fixed point iteration of $\mathbf{Z}_t$.

Algorithm 5.1 Golden Section Search for $\alpha \in [\alpha_l, \alpha_u]$ for a fixed $\mathbf{M} = \mathbf{Z}_k$.

Require: ϵ, $\eta = \frac{1+\sqrt{5}}{2}$, $\alpha_l = 0$, $\mathbf{S_{X,y}}, \mathbf{L_X}$, $\mathbf{L_y}$, $\mathbf{M} = \mathbf{Z}_k$.
Ensure: $\mathbf{Z}_{k+1} = \arg\min_{\mathbf{Z}} g(\mathbf{Z}, \mathbf{Z}_{k+1})$
$\mathbf{D}_X \leftarrow \mathbf{diag}(\mathbf{L_X})$
$\mathbf{L_M} \leftarrow 2\mathbf{M}^T\mathbf{M}$
$\alpha_u \leftarrow \lambda_{max}(\mathbf{L_M}, \mathbf{S_{X,y}})$ (Lemma A.1)
$\beta \leftarrow \alpha_u + \eta(\alpha_l - \alpha_u)$
$\delta \leftarrow \alpha_l + \eta(\alpha_u - \alpha_l)$
repeat
  $H(\beta) \leftarrow \underset{\mathbf{Z}\in\mathbb{R}^d}{\text{minimize}} \left(\mathbf{Tr}\left(\mathbf{Z}^T\mathbf{L_M}\mathbf{Z}\right) - \beta\mathbf{Tr}\left(\mathbf{Z}^T\mathbf{S_{X,y}}\mathbf{Z}\right)\right)$ (**Problem (S)**)
  $H(\delta) \leftarrow \underset{\mathbf{Z}\in\mathbb{R}^d}{\text{minimize}} \left(\mathbf{Tr}\left(\mathbf{Z}^T\mathbf{L_M}\mathbf{Z}\right) - \delta\mathbf{Tr}\left(\mathbf{Z}^T\mathbf{S_{X,y}}\mathbf{Z}\right)\right)$ (**Problem (S)**)
  if $(H(\beta) > H(\delta))$ **then**
    $\alpha_u \leftarrow \delta$, $\delta \leftarrow \beta$
    $\beta \leftarrow \alpha_u + \eta(\alpha_l - \alpha_u)$
  else
    $\alpha_l \leftarrow \beta$, $\beta \leftarrow \delta$
    $\delta \leftarrow \alpha_l + \eta(\alpha_u - \alpha_l)$
  end if
until $(|\alpha_u - \alpha_l| < \epsilon)$
$\alpha^* \leftarrow \frac{\alpha_u + \alpha_u}{2}$
$\mathbf{Z}_{k+1} \leftarrow \arg\min_{\mathbf{Z}\in\mathbb{R}^d} \left(\mathbf{Tr}\left(\mathbf{Z}^T\mathbf{L_M}\mathbf{Z}\right) - \alpha^*\mathbf{Tr}\left(\mathbf{Z}^T\mathbf{S_{X,y}}\mathbf{Z}\right)\right)$ (**Problem (S)**)
return α^*, $\mathbf{Z}_{k+1}$

Proof. From Lemma A.1 we know that, $(\gamma^2\mathbf{D_X} - \mathbf{L_M}) \succeq 0$. Hence the following would hold true for any real matrix $\mathbf{N}$,

$$\mathbf{Tr}\left((\mathbf{Z}-\mathbf{N})^T(\gamma^2\mathbf{D}_X - \mathbf{L_M})(\mathbf{Z}-\mathbf{N})\right) \geq 0$$

Rearranging the terms we get the following inequality over $\mathbf{Tr}\left(\mathbf{Z}^T\mathbf{L_M}\mathbf{Z}\right)$,

$$\begin{aligned}
\mathbf{Tr}\left(\mathbf{Z}^T\mathbf{L_M}\mathbf{Z}\right) + \mathbf{Tr}\left(\mathbf{N}^T(\gamma^2\mathbf{D}_X - \mathbf{L_M})\mathbf{Z}\right) - \mathbf{Tr}\left(\mathbf{N}^T(\gamma^2\mathbf{D}_X - \mathbf{L_M})\mathbf{N}\right) \\
\leq \mathbf{Tr}\left(\mathbf{Z}^T\gamma^2\mathbf{D}_X\mathbf{Z}\right) - \mathbf{Tr}\left(\mathbf{Z}^T(\gamma^2(\mathbf{D}_X - \mathbf{L_M})\mathbf{N}\right) \\
\mathbf{Tr}\left(\mathbf{Z}^T\mathbf{L_M}\mathbf{Z}\right) \leq \mathbf{Tr}\left(\mathbf{Z}^T\gamma^2\mathbf{D}_X\mathbf{Z}\right) - 2\mathbf{Tr}\left(\mathbf{Z}^T(\gamma^2\mathbf{D}_X - \mathbf{L_M})\mathbf{N}\right) \\
+ \mathbf{Tr}\left(\mathbf{N}^T(\gamma^2\mathbf{D}_X - \mathbf{L_M})\mathbf{N}\right) \\
= l(\mathbf{Z}, \mathbf{N})
\end{aligned}$$

If $\mathbf{N} = \mathbf{Z}$ then $l(\mathbf{Z}, \mathbf{Z}) = \mathbf{Tr}\left(\mathbf{Z}^T\mathbf{L_M}\mathbf{Z}\right)$. Hence $l(\mathbf{Z}, \mathbf{N})$ majorizes $\mathbf{Tr}\left(\mathbf{Z}^T\mathbf{L_M}\mathbf{Z}\right)$. It also follows that the surrogate function $l(\mathbf{Z}, \mathbf{N}) - \alpha\mathbf{Tr}\left(\mathbf{Z}^T\mathbf{S_{X,y}}\mathbf{Z}\right)$ majorizes our desired objective function $H(\mathbf{Z}; \alpha)$. To optimize this surrogate loss we equate its gradient to zero and rearrange the terms to obtain

$$\begin{aligned}
(\gamma^2\mathbf{D}_X - \alpha\mathbf{S_{X,y}})\mathbf{Z} &= (\gamma^2\mathbf{D}_X - \mathbf{L_M})\mathbf{N} \\
\mathbf{Z} &= (\gamma^2\mathbf{D}_X - \alpha\mathbf{S_{X,y}})^\dagger(\gamma^2\mathbf{D}_X - \mathbf{L_M})\mathbf{N},
\end{aligned}$$

which gives us the update equation $\mathbf{Z}_{t+1} = \mathbf{H}\mathbf{Z}_t$ where $\mathbf{H}$ is given by,

$$\mathbf{H} = (\gamma^2\mathbf{D}_X - \alpha\mathbf{S_{X,y}})^\dagger(\gamma^2\mathbf{D}_X - \mathbf{L_M}). \tag{5.4}$$

Hence it follows from framework of Lange (2013) that above update equation monotonically minimizes $H(\mathbf{Z};\alpha)$. □

Algorithm 5.2 summarizes the steps of an iterative Majorization-Minimization approach to solve **Problem (S)**.

Algorithm 5.2 Distance Correlation Maximization for a given α

Require: γ^2 (Theorem A.1), α, $\mathbf{M} = \mathbf{Z}_k$, $\mathbf{S}_{\mathbf{X},\mathbf{y}}$, $\mathbf{L}_\mathbf{M}$, $\mathbf{D}_\mathbf{X}$
Ensure: $H(\mathbf{Z};\alpha) = \underset{\mathbf{Z}\in\mathbb{R}^d}{\text{minimize}} \left(\mathbf{Tr}\left(\mathbf{Z}^T\mathbf{L}_\mathbf{M}\mathbf{Z}\right) - \alpha\mathbf{Tr}\left(\mathbf{Z}^T\mathbf{S}_{\mathbf{X},\mathbf{y}}\mathbf{Z}\right)\right)$

$t \leftarrow 0$
$\mathbf{Z}_t = \mathbf{Z}_k$
$H(\mathbf{Z}_t;\alpha) \leftarrow \left(\mathbf{Tr}\left(\mathbf{Z}_t^T\mathbf{L}_\mathbf{M}\mathbf{Z}_t\right) - \alpha\mathbf{Tr}\left(\mathbf{Z}_t^T\mathbf{S}_{\mathbf{X},\mathbf{y}}\mathbf{Z}_t\right)\right)$
$\mathbf{H} = \left(\gamma^2\mathbf{D}_\mathbf{X} - \alpha\mathbf{S}_{\mathbf{X},\mathbf{y}}\right)^\dagger\left(\gamma^2\mathbf{D}_\mathbf{X} - \mathbf{L}_\mathbf{M}\right)$
repeat
  $\mathbf{Z}_{t+1} = \mathbf{H}\mathbf{Z}_t$
  $H(\mathbf{Z}_{t+1};\alpha) \leftarrow \left(\mathbf{Tr}\left(\mathbf{Z}_t^T\mathbf{L}_\mathbf{M}\mathbf{Z}_t\right) - \alpha\mathbf{Tr}\left(\mathbf{Z}_t^T\mathbf{S}_{\mathbf{X},\mathbf{y}}\mathbf{Z}_t\right)\right)$
  $t \leftarrow t+1$
until $(|H(\mathbf{Z}_{t+1};\alpha) - H(\mathbf{Z}_t;\alpha)| < \epsilon)$ **or** $(t \geq T_{\max})$
$F(\alpha) \leftarrow H(\mathbf{Z}_t;\alpha)$
$\mathbf{Z}^* \leftarrow \mathbf{Z}_t$
return $F(\alpha), \mathbf{Z}^*$

6. Experiments

In this section we present experimental results that compare our proposed method with several state-of-the-art supervised dimensionality reduction techniques on a regression task.

6.1. Methodology

Methodology we use for our experiments is as follows:

(i) We run our proposed algorithm on the training set $\mathbf{X}_{\texttt{Train}}$ to learn low-dimensional features $\mathbf{Z}_{\texttt{Train}}$.
(ii) We learn the map $\psi\colon \mathbf{z} \mapsto y$ using Support Vector Regression on $\mathbf{Z}_{\texttt{Train}}$ and $\mathbf{Y}_{\texttt{Train}}$.
(iii) We learn mappings $\phi_i\colon \mathbf{x} \mapsto z_i$, $i = 1$ to d for each dimension of $\mathbf{z}$ using Support Vector Regression on $\mathbf{X}_{\texttt{Train}}$ and $\mathbf{Z}_{\texttt{Train}}$.

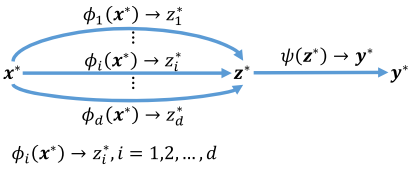


FIG 1. *Out-of-Sample prediction*

During testing/out-of-sample phase, given a test input $\mathbf{x}^*$, we use maps $\phi_i\colon \mathbf{x} \mapsto z_i$ for $i = 1$ to d and generate $\mathbf{z}^*$. We then utilize maps $\psi\colon \mathbf{z} \mapsto y$ on $\mathbf{z}^*$ to get the predicted response y^*. Figure 1 illustrates the testing phase of our methodology.

6.2. Datasets

In our results we report the Root Mean Squared (RMS) errors on five datasets from the UCI-Machine Learning Repository (Lichman, 2013) in Tables 1 to 5. We use the following datasets in our experiments.

(a) **Boston Housing** (Harrison and Rubinfeld, 1978): This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. This dataset has been used extensively throughout the vast regression literature to benchmark algorithms. The response variable to be predicted is the median value of owner-occupied homes.

(b) **Relative Location of Computed Tomography (CT) Slices** (Graf et al., 2011): This dataset consists of 385 features extracted from computed tomography (CT) images. Each CT slice is described by two histograms in polar space that are concatenated to form the final feature vector. The response variable to be predicted is the relative location of an image on the axial axis. The ground truth of responses in this dataset was constructed by manually annotating up to 10 distinct landmarks in each CT Volume with a known location. This response takes values in the range $[0, 180]$ where 0 denotes the top of the head and 180 denotes the the soles of the feet.

(c) **BlogFeedback** (Buza, 2014): This dataset originates from a set of raw HTML documents of blog posts that were crawled and processed. The task associated with this data is to predict the number of comments in the upcoming 24 hours. In order to simulate this situation, the dataset was curated by choosing a base time (in the past) and selecting the blog posts that were published at most 72 hours before the selected base date/time. Then a set of 281 features of the selected blog posts were computed from the information that was available at the basetime. The target is to predict the number of comments that the blog post received in the next 24 hours, relative to the basetime. In the training data, the base times were in the years 2010 and 2011. In the test data the base times were in February and March 2012.

(d) **Geographical Origin of Music** (Zhou et al., 2014): Instances in this dataset contain audio features extracted from 1059 wave files covering 33 countries/areas. The task associated with the data is to predict the geographical origin of music. The program MARSYAS was used to extract 68 audio features from the wave files. These were appended with 48 chromatic attributes that describe the notes of the scale bringing the total number of features to 116.

(e) **UJI Indoor Localization** (Torres-Sospedra et al., 2014): The UJIIndoorLoc is a Multi-Building Multi-Floor indoor localization database that relies on WLAN/WiFi fingerprinting technology. Automatic user localization consists of estimating the position of the user (latitude, longitude and altitude) by using an electronic device, usually a mobile phone. The task is to predict the actual longitude and latitude. The database consists of 19937 training/reference records and 1111 validation/test records. The 529 features contain the WiFi fingerprint, the coordinates where it was taken, and

other useful information. Given that this paper focusses on the setting of univariate responses, we only aim to predict the 'Longitude'.

6.3. Results

We perform five-fold cross validation on each of these datasets and report the average Root Mean Square (RMS) error on the hold-out test sets. Tables 1 to 5 present the cross-validated RMS error of our proposed method (DisCoMax), and six other supervised dimensionality reduction techniques namely; LSDR (Suzuki and Sugiyama, 2013), gKDR (Fukumizu and Leng, 2014), SCA (Yamada et al., 2011), LAD (Cook and Forzani, 2009), SAVE (Shao et al., 2009) and (Shao et al., 2007) and SIR (Li, 1991).

In case of DisCoMax, we use the methodology described in sub-section 6.1. For other methods we used in our evaluation, these techniques generate explicit maps to obtain the low-dimensional representations. As in the case of the methodology for DisCoMax, we use these explicit maps and Support Vector Regression (with a RBF kernel) to generate cross-validated RMS errors on the responses.

We fix the folds used across the seven techniques presented within each of the tables (Tables 1 to 5). We also compute RMS errors for increasing dimensions $d = 3, 5, 7, 9$ and 11. We note a significant improvement in the predictive performance of DisCoMax learnt features across all above mentioned cases of chosen dimensionality d. We also note that the predictive performance (smaller error) increases at a slower rate as we increase dimensionality of learnt features. This experimental setup mimics the setup of Fukumizu and Leng (2014) and hence at the moment, we just choose d such that the cross-validation error does not decrease substantially (although a subjective choice) with any further increases. We also believe that a better choice of estimating an optimal d from the data prior to running the algorithm, would be further helpful.

For baseline comparison purposes, in case of the Boston Housing dataset, we observe a RMS error of 0.1719 using Support Vector Regression without any dimensionality reduction ($d = 13$). This when compared to DisCoMax RMS errors which ranged between 0.1559 ($d = 3$) and 0.1297 ($d = 11$) always did worse. We bold errors for DisCoMax for cases where errors were significantly better when compared with their corresponding standard deviations taken into account.

We show in Figure 2 the effect of increasing distance correlations by plotting the iteratively learnt $\mathbf{Z}$ and fixed response $\mathbf{y}$ for different iterations of 0,500 and 1000 respectively. In the first row, we compare the response variable and the learnt feature variable $\mathbf{Z_{zn}}$ obtained by applying our technique on a feature variable called "zn" in the original feature matrix $\mathbf{X}$ of the Boston Housing dataset. We present the results for the same for the variables of "rm" and "indus" in the next two rows of this Figure.

Similarly, in Figure 3 we plot the iteratively learnt $\mathbf{Z}$ and corresponding feature variable $\mathbf{X}$ for different iterations of 0,500 and 1000 for the variables of "zn", "rm" and "indus" respectively. We show that the distance correlations decrease

TABLE 1
Boston Housing (Harrison and Rubinfeld, 1978): U.S Census Service concerning housing in the area of Boston Mass. To predict median value of owner-occupied homes. Baseline results SVR RMS error 0.1719.

Method/dimension	3	5	7	9	11
DisCoMax	**0.1559**	**0.1493**	**0.1327**	**0.1311**	**0.1297**
LSDR (Suzuki and Sugiyama, 2013)	0.1978	0.1963	0.1892	0.1886	0.1873
gKDR (Fukumizu and Leng, 2014)	0.1997	0.1813	0.1762	0.1738	0.1719
SCA (Yamada et al., 2011)	0.1875	0.1796	0.1708	0.1637	0.1602
LAD (Cook and Forzani, 2009)	0.2019	0.1964	0.1932	0.1917	0.1903
SAVE (Shao et al., 2009)	0.2045	0.1983	0.1967	0.1952	0.1947
SIR (Li, 1991)	0.2261	0.2193	0.2086	0.2076	0.2068

TABLE 2
Geographical Origin of Music (Graf et al., 2011): The input contains audio features extracted from 1059 wave files covering 33 countries/areas. The task associated with the data is to predict the geographical origin of music.

Method/d	3	5	7	9	11
DisCoMax	**19.19**	**18.67**	**18.14**	**17.94**	**17.81**
LSDR (Suzuki and Sugiyama, 2013)	23.63	22.31	22.09	21.93	21.82
gKDR (Fukumizu and Leng, 2014)	24.06	23.39	22.76	22.52	22.50
SCA (Yamada et al., 2011)	23.17	24.96	24.21	23.34	23.06
LAD (Cook and Forzani, 2009)	26.74	25.57	24.39	24.26	24.20
SAVE (Shao et al., 2009)	28.18	27.82	27.62	27.53	27.50
SIR (Li, 1991)	29.92	29.46	29.18	28.86	28.63

TABLE 3
BlogFeedback (Buza, 2014): This data contains features computed from raw HTML documents of blog posts. The task associated with this data is to predict the number of comments in the upcoming 24 hours.

Method/d	3	5	7	9	11
DisCoMax	**25.82**	**24.69**	**24.33**	**23.90**	**23.62**
LSDR (Suzuki and Sugiyama, 2013)	30.36	28.16	27.39	27.24	27.18
gKDR (Fukumizu and Leng, 2014)	29.72	27.62	27.29	26.91	26.81
SCA (Yamada et al., 2011)	28.53	27.31	26.60	26.32	26.30
LAD (Cook and Forzani, 2009)	30.42	30.39	30.20	30.04	29.99
SAVE (Shao et al., 2009)	31.93	31.27	30.72	30.53	30.31
SIR (Li, 1991)	33.63	32.65	31.39	31.16	30.83

at a relatively slower rate in comparison to the Figure 2. This hints towards the learnt $\mathbf{Z}$ being a *compressed* representation of $\mathbf{X}$ while also maintaining a higher distance correlation with $\mathbf{Y}$.

TABLE 4

Relative location of CT slices (Zhou et al., 2014): Dataset consists of 385 features extracted from CT images. Features are concatenation of two histograms in polar space. The response variable is the relative location of an image on the axial axis.

Method/d	3	5	7	9	11
DisCoMax	**12.29**	**11.11**	**10.19**	**9.73**	**9.66**
LSDR (Suzuki and Sugiyama, 2013)	14.38	13.14	12.87	12.73	12.69
gKDR (Fukumizu and Leng, 2014)	13.65	12.86	12.67	12.35	12.05
SCA (Yamada et al., 2011)	14.19	13.64	12.94	12.12	11.73
LAD (Cook and Forzani, 2009)	17.70	17.62	17.34	17.15	16.89
SAVE (Shao et al., 2009)	19.32	18.74	18.62	17.76	17.21
SIR (Li, 1991)	21.53	21.23	20.97	20.77	20.64

TABLE 5

UJI Indoor Localization (Torres-Sospedra et al., 2014): Multi-Building Multi-Floor indoor localization database. The task is to predict the actual longitude and latitude. The 529 attributes contain the WiFi fingerprint, the coordinates where it was taken. The database consists of around 20ktraining/reference records and 11k validation/test records.

Method/d	3	5	7	9	11
DisCoMax	**12.28**	**11.10**	**10.19**	**9.73**	**9.65**
LSDR (Suzuki and Sugiyama, 2013)	14.38	13.14	12.86	12.73	12.69
gKDR (Fukumizu and Leng, 2014)	13.65	12.86	12.67	12.34	12.05
SCA (Yamada et al., 2011)	14.18	13.63	12.94	12.12	11.73
LAD (Cook and Forzani, 2009)	17.69	17.62	17.34	17.15	16.89
SAVE (Shao et al., 2009)	19.32	18.74	18.61	17.75	17.20
SIR (Li, 1991)	21.53	21.23	20.97	20.77	20.63

7. Discussion

In this section, we discuss effects of choice of α in the optimization of **Problem (S)** (Algorithm 5.2). We also experimentally show results on the iterative optimization of **Problem (P)** using Algorithm 4.1 which optimizes a lower bound in **Problem (Q)**. We use the Boston Housing dataset for our analysis.

Figures 4a and 4b show gradual increase in sample distance correlations $\hat{\rho}(\mathbf{X}, \mathbf{Z}_t)$ (Blue) and $\hat{\rho}(\mathbf{Z}_t, \mathbf{y})$ (Red) with respect to the number of fixed point t for two different choices of $\alpha = 6 \times 10^4$ and $\alpha = 70 \times 10^4$. We clearly observe that the choice of α has a strong effect on rate of increase/decrease of individual distance correlations $\hat{\rho}^2(\mathbf{X}, \mathbf{Z}_t)$ and $\hat{\rho}^2(\mathbf{Z}_t, \mathbf{y})$ as iterations progress. This is because the α value positively weighs the term $\mathbf{Tr}\left(\mathbf{Z}^T \mathbf{S}_{\mathbf{X},\mathbf{y}} \mathbf{Z}\right)$ over $\mathbf{Tr}\left(\mathbf{Z}^T \mathbf{L}_{\mathbf{M}} \mathbf{Z}\right)$ in **Problem (S)**. Figure 4c shows the rate of change of objective function $f(\mathbf{Z})$ with respect to the fixed point iterations t for two choices of α. The figure clearly shows the slower (faster) rate of increase of $f(\mathbf{Z})$ for smaller (larger) α.

Figure 5a and 5b repectively show the overall growth of distance correlations $(\hat{\rho}(\mathbf{X}, \mathbf{Z}),\ \hat{\rho}(\mathbf{Z}, \mathbf{y}))$ and $f(\mathbf{Z})$, with respect to the fixed point iterations (t), for

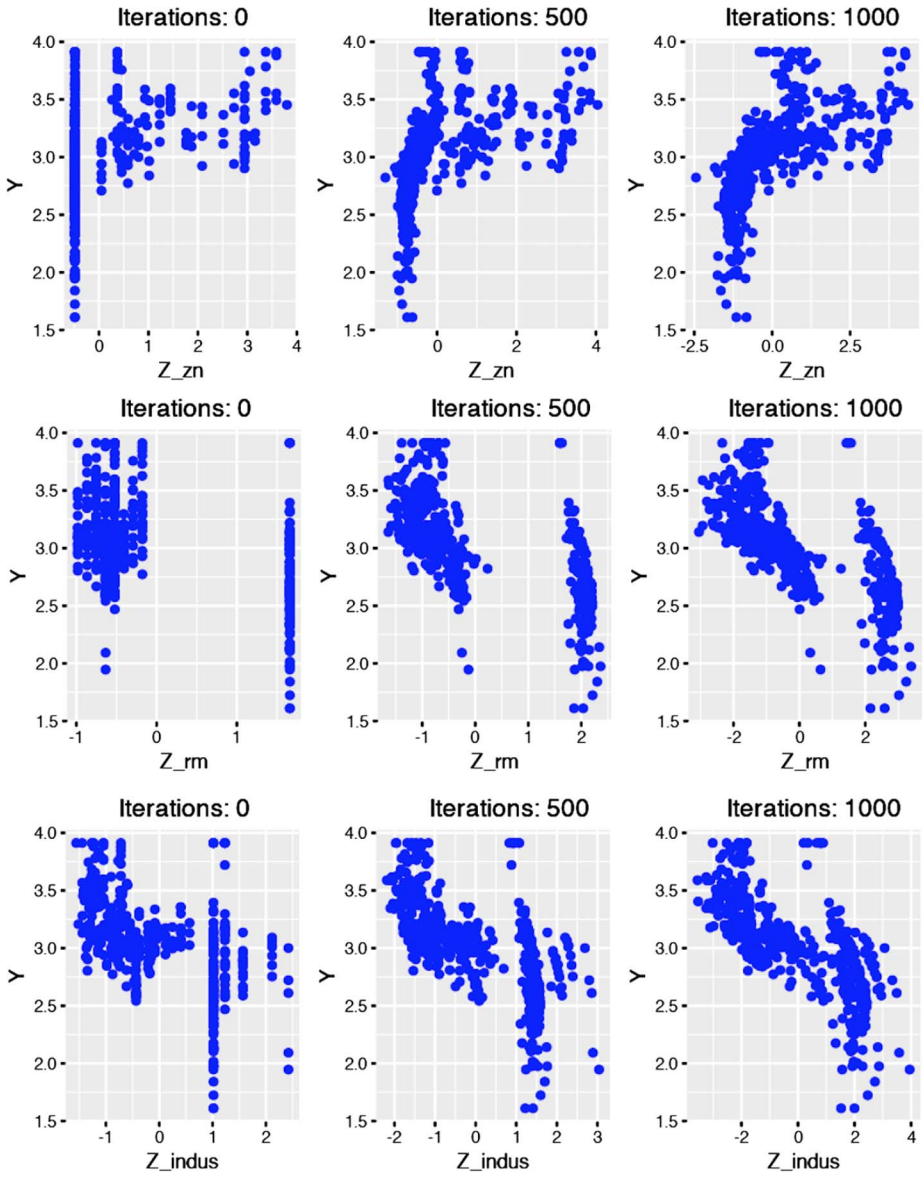


FIG 2. *Learnt representations Z corresponding to various variables in X vs. response Y after 0, 500 & 1000 iterations of our algorithm.*

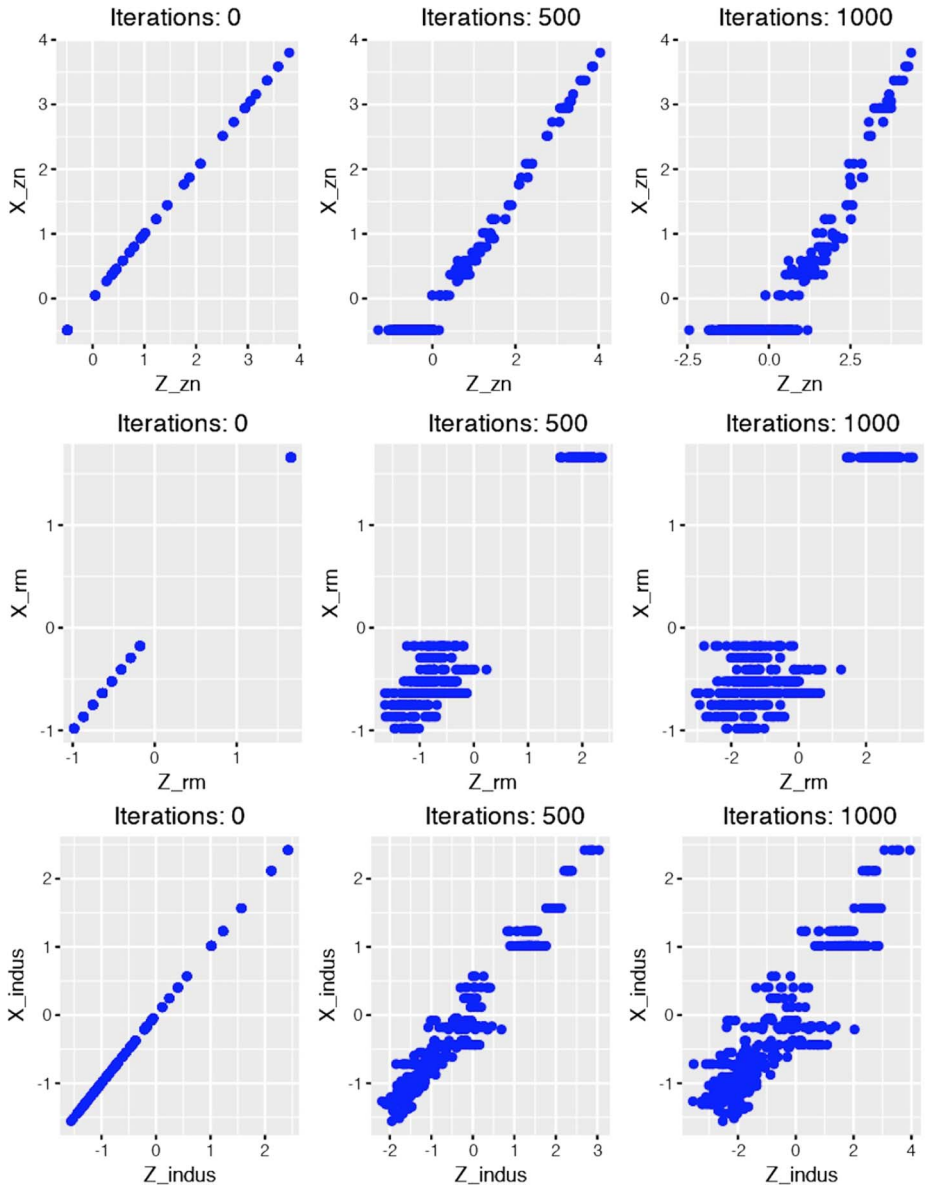
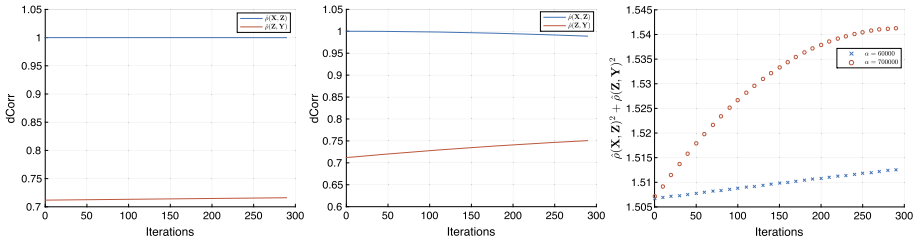


FIG 3. *Learnt representations Z corresponding to various variables in X vs. the original X variable after 0, 500 & 1000 iterations of our algorithm.*



(a) $\hat{\rho}(\mathbf{X}, \mathbf{Z})$ (Blue) and $\hat{\rho}(\mathbf{Z}, \mathbf{y})$ (Red) vs. fixed point iterations (t) for $\alpha = 6 \times 10^4$

(b) $\hat{\rho}(\mathbf{X}, \mathbf{Z})$ (Blue) and $\hat{\rho}(\mathbf{Z}, \mathbf{y})$ (Red) vs. fixed point iterations (t) for $\alpha = 70 \times 10^4$

(c) $f(\mathbf{Z})$ vs. fixed point iterations for $\alpha = 6 \times 10^4$ (Red) for $\alpha = 70 \times 10^4$ (Blue).

FIG 4. *Effect of α values on growth of the proposed objective in Algorithm 5.2 the figures show slower (faster) growth of distance correlations for smaller (larger) α.*

$\alpha^* = 800 \times 10^4$. We periodically observe a sharp increases in $f(\mathbf{Z})$ and distance correlations after each DisCoMax subproblem of 220 fixed point iterations. The figures show four such G-MM iterations of Algorithm 4.1. These sharp increases are due to the resubstitution of $\mathbf{M} = \mathbf{Z}_k$ in Step 2 of Algorithm 4.1. This clearly shows us that we are able to iteratively maximize the proposed objective in **Problem (P)**.

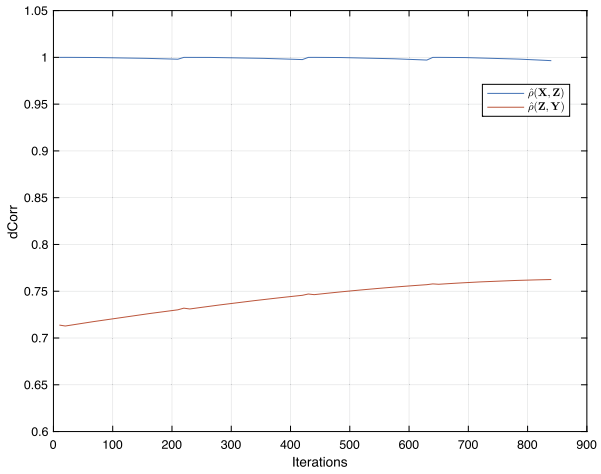
8. Conclusion

In our work, we proposed a novel method to perform supervised dimensionality reduction. Our method aims to maximize an objective based on a statistical measure of dependence called statistical distance correlation. Our proposed method does not necessarily constrain the dimension reduction projection to be linear. We also propose a novel algorithm to optimize our proposed objective using the Generalized Minorization-Maximization approach of Parizi et al. (2015). Finally, we show a superior empirical performance of our method on several regression problems in comparison to existing state-of-the-art methods.

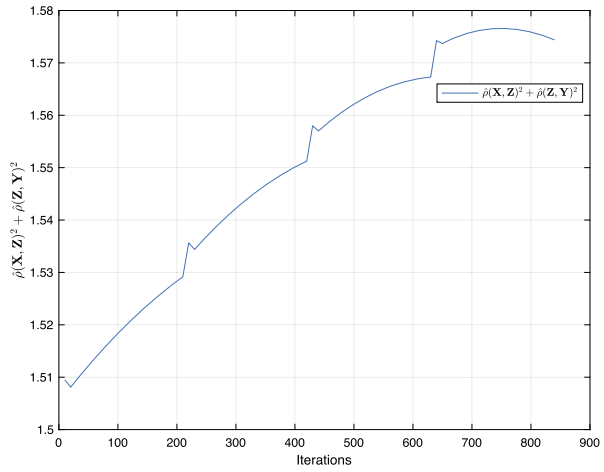
For future work, we aim to extend our framework to handle multivariate responses $\mathbf{y} \in \mathbb{R}^q$, as distance correlation is applicable to variables with arbitrary dimensions. Our proposed approach is practically applicable on relatively small datasets, as it involves repeatedly solving multiple optimization subproblems. So we aim to to simplyfy this approach so that it is tractable for larger size (several thousands of examples) datasets. In our work, we currently tackle the out-of-sample issue by learning mutiple Support Vector Regressions, one for each dimension of $\mathbf{z}$. We plan to extend our framework so as to learn explicit out-of-sample mappings from $\mathbf{x}$ to $\mathbf{z}$.

Appendix A: Spectral radius of the fixed point iterate $T(\mathbf{Z}_t)$

To prove Lemma A.4, required for proving convergence in Theorem 4.1, we need to show that the spectral radius $\lambda_{max}(\mathbf{H}) < 1$. We show this in Theorem A.3



(a) $\hat{\rho}(\mathbf{X},\mathbf{Z})$ (Blue) and $\hat{\rho}(\mathbf{Z},\mathbf{y})$ (Red) vs overall Iterations.



(b) $f(\mathbf{Z}) = \hat{\rho}(\mathbf{X},\mathbf{Z})^2 + \hat{\rho}(\mathbf{Z},\mathbf{y})^2$ vs overall Iterations.

FIG 5. *Overall gradual increase in* $f(\mathbf{Z})$ *(Figure 5a) and distance correlations (Figure 5b) for* $\alpha^* = 800\times10^4$. *Plots show increase in both for each DisCoMax subproblem of (Algorithm 5.2) and four outer G-MM iterations of Algorithm 4.1*

and proceed to prove it by first by proving two required lemmas below.

Lemma A.1. *For any choice of* $\gamma^2 > \lambda_{max}(\mathbf{D_X},\mathbf{L_M})$ *and* $\mathbf{P}\colon = \left(\gamma^2\mathbf{D_X} - \mathbf{L_M}\right)$, *we have* $\mathbf{P} \succeq 0$.

Proof. To show $\mathbf{z}^T(\gamma^2\mathbf{D_X} - \mathbf{L_M})\mathbf{z} \geq 0$ for all $\mathbf{z}$, we require that $\gamma^2 \geq \frac{\mathbf{z}^T\mathbf{L_M}\mathbf{z}}{\mathbf{z}^T\mathbf{D_X}\mathbf{z}}$ for all $\mathbf{z}$. This is always true for all values of $\gamma^2 \geq \lambda_{max}(\mathbf{D_X},\mathbf{L_M})$. □

Lemma A.2. *If* $0 = \alpha_l \leq \alpha \leq \alpha_u = \lambda_{min}(\mathbf{L_M}, \mathbf{S_{X,y}})$ *and* $\mathbf{Q} := (\mathbf{L_M} - \alpha \mathbf{S_{X,y}})$, *then we have* $\mathbf{Q} \succeq 0$.

Proof. To show $\mathbf{z}^T(\mathbf{L_M} - \alpha \mathbf{S_{X,y}})\mathbf{z} \geq 0$ for all $\mathbf{z}$, we require that $\alpha \leq \frac{\mathbf{z}^T \mathbf{L_M z}}{\mathbf{z}^T \mathbf{S_{X,y}} \mathbf{z}}$ for all $\mathbf{z}$. This is always true if all values of $\alpha \leq \min_{\mathbf{Z}} \frac{\mathbf{z}^T \mathbf{L_M z}}{\mathbf{z}^T \mathbf{S_{X,y}} \mathbf{z}} = \lambda_{min}(\mathbf{L_M}, \mathbf{S_{X,y}})$ which is true by our choice of α. □

We now utilize the above to results to prove $\lambda_{max}(\mathbf{H}) \leq 1$ about the fixed point iterate $\mathbf{Z}_{t+1} = \mathbf{H}\mathbf{Z}_t$.

Theorem A.3. *For the update equation* $\mathbf{Z}_{t+1} = \mathbf{H}\mathbf{Z}_t$ *with*

$$\mathbf{H} = \left(\gamma^2 \mathbf{D_X} - \alpha \mathbf{S_{X,y}}\right)^{\dagger} \left(\gamma^2 \mathbf{D_X} - \mathbf{L_M}\right),$$

we have $\lambda_{max}(\mathbf{H}) \leq 1$.

Proof. The update equation looks as follows

$$\mathbf{Z}_{t+1} = \left(\gamma^2 \mathbf{D_X} - \alpha \mathbf{S_{X,y}}\right)^{\dagger} \left(\gamma^2 \mathbf{D_X} - \mathbf{L_M}\right) \mathbf{Z}_t.$$

For sake of simplicity assume $\mathbf{P} = \left(\gamma^2 \mathbf{D_X} - \mathbf{L_M}\right)$ and $\mathbf{Q} = (\mathbf{L_M} - \alpha \mathbf{S_{X,y}})$.

$$\mathbf{Z}_{t+1} = (\mathbf{P} + \mathbf{Q})^{-1} \mathbf{P}\mathbf{Z}_t$$

Using the Woodbury matrix identity $(\mathbf{A} + \mathbf{UBV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{B}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}$, and setting $\mathbf{U} = \mathbf{I}$ and $\mathbf{V} = \mathbf{I}$, we get, $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{B}^{-1} + \mathbf{A}^{-1})^{-1}\mathbf{A}^{-1}$. Applying this to the previous equation we get

$$\begin{aligned}\mathbf{Z}_{t+1} &= (\mathbf{P}^{-1} - \mathbf{P}^{-1}(\mathbf{P}^{-1} + \mathbf{Q}^{-1})^{-1}\mathbf{P}^{-1})\mathbf{P}\mathbf{Z}_t = \mathbf{I} - \mathbf{P}^{-1}(\mathbf{P}^{-1} + \mathbf{Q}^{-1})^{-1}\mathbf{Z}_t \\ &= \mathbf{I} - \mathbf{P}^{-1}\left((\mathbf{P}^{-1} + \mathbf{Q}^{-1})^{-1}\mathbf{Q}^{-1}\right)\mathbf{Q}\mathbf{Z}_t\end{aligned}$$

Using the positive definite identity $(\mathbf{P}^{-1} + \mathbf{B}^T\mathbf{Q}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Q}^{-1} = \mathbf{P}\mathbf{B}^T(\mathbf{B}\mathbf{P}\mathbf{B}^T + \mathbf{Q})^{-1}$ for $\mathbf{B} = \mathbf{I}$ we get, $(\mathbf{P}^{-1} + \mathbf{Q}^{-1})^{-1}\mathbf{Q}^{-1} = \mathbf{P}(\mathbf{P} + \mathbf{Q})^{-1}$, which simplifies the term in the brackets as,

$$\mathbf{Z}_{t+1} = \mathbf{I} - \mathbf{P}^{-1}\left(\mathbf{P}(\mathbf{P} + \mathbf{Q})^{-1}\right)\mathbf{Q}\mathbf{Z}_t = \mathbf{I} - (\mathbf{P} + \mathbf{Q})^{-1}\mathbf{Q}\mathbf{Z}_t$$

If we compare the above equation with a the general update equation from Zhang et al. (2000), which is of the form

$$T(\mathbf{Z}_{t+1}) = \mathbf{Z}_t - \beta(\mathbf{Z}_t)\mathbf{B}(\mathbf{Z}_t)^{-1}\nabla f(\mathbf{Z}_t)$$

where $\nabla f(\mathbf{Z}_t)$ is the gradient of the objective function $f(\mathbf{Z})$ we get,

$$\beta(\mathbf{Z}_t) = \frac{1}{2}, \qquad \mathbf{B}(\mathbf{Z}_t) = \mathbf{P} + \mathbf{Q}, \qquad \nabla f(\mathbf{Z}_t) = 2\mathbf{Q}\mathbf{Z}_t$$

Now from Theorem A.1 we conclude that $\mathbf{B}(\mathbf{Z}) \succeq 0$, We also check the following condition from Zhang et al. (2000) that

$$0 \preceq \nabla^2 f(\mathbf{Z}) \preceq \frac{2\mathbf{B}}{\beta}.$$

or equivalently, as in our case $0 \preceq 2\mathbf{Q} \preceq 4(\mathbf{Q} + \mathbf{P})$, which is indeed true. Hence it follows that $\lambda_{max}(T'(\mathbf{Z})) \leq 1$ which implies $\lambda_{max}(\mathbf{H}) \leq 1$. □

We now proceed to show that at end of every $(t+1)$ fixed point iterations we have $\mathbf{Tr}\left(\mathbf{Z}_{t+1}^T\mathbf{L}_{\mathbf{Z}_{t+1}}\mathbf{Z}_{t+1}\right) \leq \mathbf{Tr}\left(\mathbf{Z}_{t+1}\mathbf{L}_{\mathbf{Z}_0}\mathbf{Z}_{t+1}\right)$.

Lemma A.4. *For fixed point iteration* $\mathbf{Z}_{t+1} = \mathbf{H}\mathbf{Z}_t$ *for optimization of* $\mathbf{Z}_{k+1} = \arg\max_{\mathbf{Z}} g(\mathbf{Z}, \mathbf{Z}_k)$*, we have,* $\mathbf{Tr}\left(\mathbf{Z}_{k+1}^T\mathbf{L}_{\mathbf{Z}_{k+1}}\mathbf{Z}_{k+1}\right) \leq \mathbf{Tr}\left(\mathbf{Z}_{k+1}\mathbf{L}_{\mathbf{Z}_k}\mathbf{Z}_{k+1}\right)$.

Proof. Laplacian for a weighted adjacency matrix $\mathbf{W}$ (with self loops) is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$ where $\mathbf{D}$ is a diagonal degree matrix with diagonal elements $[\mathbf{D}]_{i,i} = \sum_j [\mathbf{W}]_{i,j}$ and zero off-diagonal entries (Chung, 1997). For adjacency matrix $\widehat{\mathbf{E}}_{\mathbf{Z}}$ we have $\widehat{\mathbf{E}}_{\mathbf{Z}} = \mathbf{J}\mathbf{E}_{\mathbf{Z}}\mathbf{J} = -2\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T$ (Torgerson, 1952). We have Laplacian as $\mathbf{L}_{\mathbf{Z}} = \mathbf{D}_{\mathbf{Z}} - \widehat{\mathbf{E}}_{\mathbf{Z}}$ with $\mathbf{D}_{\mathbf{Z}} = 0$. This gives us for $\mathbf{Z}_{t+1}$ the Laplacian $\mathbf{L}_{\mathbf{Z}_{t+1}} = 2\mathbf{Z}_{t+1}\mathbf{Z}_{t+1}^T$. It also follows from the fact that since we choose our initialization $\mathbf{Z}_0$ as column-centered matrix, and $\mathbf{Z}_{t+1} = \mathbf{H}\mathbf{Z}_t$ are also successively column-centered for all $t > 0$. Hence, $\mathbf{L}_{\mathbf{Z}_{t+1}} = 2\widehat{\mathbf{Z}}_{t+1}\widehat{\mathbf{Z}}_{t+1}^T$. Now substituting $\mathbf{Z}_{t+1} = \mathbf{H}\mathbf{Z}_t$ in Laplacian equation $\mathbf{L}_{\mathbf{Z}_{t+1}}$ we get,

$$\mathbf{L}_{\mathbf{Z}_{t+1}} = 2(\mathbf{H}\mathbf{Z}_t)(\mathbf{H}\mathbf{Z}_t)^T = 2\mathbf{H}\mathbf{Z}_t\mathbf{Z}_t^T\mathbf{H}^T = \mathbf{H}\mathbf{L}_{\mathbf{Z}_t}\mathbf{H}^T. \tag{A.1}$$

Substituting above equation into right hand side of the statement to be proved gives us,

$$\mathbf{Tr}\left(\mathbf{Z}_{t+1}^T\mathbf{L}_{\mathbf{Z}_{t+1}}\mathbf{Z}_{t+1}\right) = \mathbf{Tr}\left(\mathbf{Z}_{t+1}^T\mathbf{H}\mathbf{L}_{\mathbf{Z}_t}\mathbf{H}^T\mathbf{Z}_{t+1}\right).$$

Substituting eigen decomposition of $\mathbf{H} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ where $\mathbf{\Lambda}$ is a diagonal eigenvalues matrix with values less than one (Theorem A.3) we get,

$$\mathbf{Tr}\left(\mathbf{Z}_{t+1}\mathbf{L}_{\mathbf{Z}_{t+1}}\mathbf{Z}_{t+1}\right) = \mathbf{Tr}\left(\mathbf{Z}_{t+1}^T(\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T)\mathbf{L}_{\mathbf{Z}_t}(\mathbf{Q}^T\mathbf{\Lambda}\mathbf{Q})\mathbf{Z}_{t+1}\right).$$

For $\mathbf{\Lambda} = \mathbf{I}$ (identity matrix) gives us,

$$\begin{aligned}\mathbf{Tr}\left(\mathbf{Z}_{t+1}\mathbf{L}_{\mathbf{Z}_{t+1}}\mathbf{Z}_{t+1}\right) &\leq \mathbf{Tr}\left(\mathbf{Z}_{t+1}^T(\mathbf{Q}\mathbf{I}\mathbf{Q}^T)\mathbf{L}_{\mathbf{Z}_t}(\mathbf{Q}^T\mathbf{I}\mathbf{Q})\mathbf{Z}_{t+1}\right)\\ &\leq \mathbf{Tr}\left(\mathbf{Z}_{t+1}^T\mathbf{L}_{\mathbf{Z}_t}\mathbf{Z}_{t+1}\right).\end{aligned}$$

Repeating the above process until $t = 0$ we get $\mathbf{Tr}\left(\mathbf{Z}_{t+1}\mathbf{L}_{\mathbf{Z}_{t+1}}\mathbf{Z}_{t+1}\right) \leq \mathbf{Tr}\left(\mathbf{Z}_{t+1}^T\mathbf{L}_{\mathbf{Z}_0}\mathbf{Z}_{t+1}\right)$. Now, for the initialisation $\mathbf{Z}_t = \mathbf{Z}_k$ at $t = 0$, and given that $\mathbf{Z}_{k+1} = \arg\max_{\mathbf{Z}} g(\mathbf{Z}, \mathbf{Z}_k)$ we have,

$$\mathbf{Tr}\left(\mathbf{Z}_{k+1}\mathbf{L}_{\mathbf{Z}_{k+1}}\mathbf{Z}_{k+1}\right) \leq \mathbf{Tr}\left(\mathbf{Z}_{k+1}^T\mathbf{L}_{\mathbf{Z}_k}\mathbf{Z}_{k+1}\right). \qquad \square$$

Lemma A.4 above allows us to show the following corollary:

Corollary 1. *For fixed point iteration* $\mathbf{Z}_{t+1} = \mathbf{H}\mathbf{Z}_t$ *optimization of* $\mathbf{Z}_{k+1} = \arg\max_{\mathbf{Z}} g(\mathbf{Z}, \mathbf{Z}_k)$*, we have* $\mathbf{Tr}\left(\mathbf{Z}_{k+1}\mathbf{L}_{\mathbf{Z}_{k+1}}\mathbf{Z}_{k+1}\right) \leq \mathbf{Tr}\left(\mathbf{Z}_k^T\mathbf{L}_{\mathbf{Z}_k}\mathbf{Z}_k\right)$.

Proof. From Lemma A.4 we have

$$\mathbf{Tr}\left(\mathbf{Z}_{k+1}\mathbf{L}_{\mathbf{Z}_{k+1}}\mathbf{Z}_{k+1}\right) \leq \mathbf{Tr}\left(\mathbf{Z}_{k+1}^T\mathbf{L}_{\mathbf{Z}_k}\mathbf{Z}_{k+1}\right) \leq \mathbf{Tr}\left(\mathbf{Z}_k^T\mathbf{H}^T\mathbf{L}_{\mathbf{Z}_k}\mathbf{H}\mathbf{Z}_k\right)$$

Following approach similar to proof of Lemma A.4 above by substituting eigen decomposition of $\mathbf{H} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ into equation above we get,

$$\mathbf{Tr}\left(\mathbf{Z}_{k+1}\mathbf{L}_{\mathbf{Z}_{k+1}}\mathbf{Z}_{k+1}\right) \leq \mathbf{Tr}\left(\mathbf{Z}_k^T((\mathbf{Q}^T\mathbf{I}\mathbf{Q})^T)\mathbf{L}_{\mathbf{Z}_k}(\mathbf{Q}^T\mathbf{I}\mathbf{Q})\mathbf{Z}_k\right) \leq \mathbf{Tr}\left(\mathbf{Z}_k^T\mathbf{L}_{\mathbf{Z}_k}\mathbf{Z}_k\right)$$

$\square$

References

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276.

Berrendero José R, C. A. and Torrecilla, J. L. (2014). Variable selection in functional data classification: a maxima-hunting proposal. *Statistica Sinica*.

Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Buza, K. (2014). Feedback prediction for blogs. In *Data analysis, machine learning and knowledge discovery*, pages 145–152. Springer.

Chechik Gal, Globerson Amir, T. N. and Yair, W. (2005). Information bottleneck for gaussian variables. *Journal of Machine Learning Research*. MR2249818

Chung, F. (1997). Lecture notes on spectral graph theory. *Providence, RI: AMS Publications*.

Cook, R. and Forzani, L. (2009). Likelihood based sufficient dimension reduction. *Journal of the American Statistical Association*, 104:197–208. MR2504373

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91(435):983–992.

Dinkelbach, W. (1967). On nonlinear fractional programming. *Management Science. Journal of the Institute of Management Science. Application and Theory Series*, 13(7):492–498.

Fukumizu, K. and Leng, C. (2014). Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Association*, 109(505):359–370. MR3180569

Graf, F., Kriegel, H.-P., Schubert, M., Pölsterl, S., and Cavallaro, A. (2011). 2d image registration in CT images using radial image descriptors. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*, pages 607–614. Springer.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102.

Kiefer, J. (1953). Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4(3):502–506. MR0055639

Kong, J., Wang, S., and Wahba, G. (2015). Using distance covariance for improved variable selection with application to learning genetic risk models. *Statistics in medicine*, 34(10):1708–1720.

Lange, K. (2013). The MM algorithm. In *Optimization*, volume 95 of *Springer Texts in Statistics*, pages 185–219. Springer New York. MR3052733

Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation

learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Lichman, M. (2013). UCI machine learning repository.

Lue, H. H. (2009). Sliced inverse regression for multivariate response regression. *Journal of Statistical Planning and Inference*, 139:2656–2664.

Nishimori, Y. and Akaho, S. (2005). Learning algorithms utilizing quasi-geodesic flows on the stiefel manifold. *Neurocomputing*, 67:106–135.

Noam, S. (2002). The information bottleneck: Theory and applications. *Ph.D. Thesis: Hebrew University of Jerusalem*.

Parizi, S. N., He, K., Sclaroff, S., and Felzenszwalb, P. (2015). Generalized majorization-minimization. *arXiv preprint arXiv:1506.07613*.

Ravid, S.-Z. and Naftali, T. (2017). Opening the black box of deep neural networks via information. *https://arxiv.org/abs/1703.00810*.

Schaible, S. (1976). Minimization of ratios. *Journal of Optimization Theory and Applications*, 19(2):347–352.

Shao, Y., Cook, R., and Weisberg, S. (2007). Marginal tests with sliced average variance estimation. *Biometrika*, 94:285–296.

Shao, Y., Cook, R., and Weisberg, S. (2009). Partial central subspace and sliced average variance estimation. *Journal of Statistical Planning and Inference*, 139:952–961.

Sheng, W. and Yin, X. (2016). Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics*, 25(1):91–104.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press, New York, NY, USA, 1st edition. MR2895762

Suzuki, T. and Sugiyama, M. (2013). Sufficient dimension reduction via squared-loss mutual information estimation. *Neural computation*, 25(3):725–758.

Székely, G. J. and Rizzo, M. L. (2012). On the uniqueness of distance covariance. *Statistics & Probability Letters*, 82(12):2278–2282. MR2979766

Székely, G. J. and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213. MR3053543

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794.

Székely, G. J., Rizzo, M. L., et al. (2009). Brownian distance covariance. *Annals of Applied Statistics*, 3(4):1236–1265.

Szekely, J. G., Rizzo, L. M., and Bakirov, K. N. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35:2769–2794.

Szretter, M. E. and Yohai, V. J. (2009). The sliced inverse regression algorithm as a maximum likelihood procedure. *Journal of Statistical Planning and Inference*, 139:3570–3578.

Tishby Naftali, P. F. C. and William, B. (1999). The information bottleneck method. *The 37th annual Allerton Conference on Communication, Control, and Computing*.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419. MR0054219

Torres-Sospedra, J., Montoliu, R., Martınez-Usó, A., Avariento, J. P., Arnau, T. J., Benedito-Bordonau, M., and Huerta, J. (2014). Ujiindoorloc: A new multi-building and multi-floor database for wlan fingerprint-based indoor localization problems. In *Proceedings of the fifth conference on indoor positioning and indoor navigation*.

Vapnik, V., Braga, I., and Izmailov, R. (2015). Constructive setting for problems of density ratio estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8(3):137–146. MR3353279

Xin Chen, D. C. and Zou, C. (2015). Diagnostic studies in sufficient dimension reduction. *Biometrika*, 102(3):545–558.

Yamada, M., Niu, G., Takagi, J., and Sugiyama, M. (2011). Sufficient component analysis for supervised dimension reduction. *arXiv preprint arXiv:1103.4998*.

Zhang, A. (2008). *Quadratic fractional programming problems with quadratic constraints*. PhD thesis, Kyoto University.

Zhang, Y., Tapia, R., and Velazquez, L. (2000). On convergence of minimization methods: attraction, repulsion, and selection. *Journal of Optimization Theory and Applications*, 107(3):529–546.

Zhou, F., Claire, Q., and King, R. D. (2014). Predicting the geographical origin of music. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 1115–1120. IEEE.