LEARNING NON-LINEAR TRANSFORM WITH DISCRIMINATIVE AND MINIMUM INFORMATION LOSS PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper proposes a novel approach for learning discriminative and sparse representations. It consists of utilizing two different models. A predefined number of non-linear transform models are used in the learning stage, and one sparsifying transform model is used at test time. The non-linear transform models have discriminative and minimum information loss priors. A novel measure related to the discriminative prior is proposed and defined on the support intersection for the transform representations. The minimum information loss prior is expressed as a constraint on the conditioning and the expected coherence of the transform matrix. An equivalence between the non-linear models and the sparsifying model is shown only when the measure that is used to define the discriminative prior goes to zero. An approximation of the measure used in the discriminative prior is addressed, connecting it to a similarity concentration. To quantify the discriminative properties of the transform representation, we introduce another measure and present its bounds. Reflecting the discriminative quality of the transform representation we name it as discrimination power.

To support and validate the theoretical analysis a practical learning algorithm is presented. We evaluate the advantages and the potential of the proposed algorithm by a computer simulation. A favorable performance is shown considering the execution time, the quality of the representation, measured by the discrimination power and the recognition accuracy in comparison with the state-of-the-art methods of the same category.

1 INTRODUCTION

Learning a transform that provides sparse and discriminative representation is an active domain of research in many areas, it includes but is not limited to data processing, pattern recognition, image processing, language modeling, text analysis, gene separation, etc. A class of algorithms proposed by (Kreutz-Delgado et al., 2003; Mairal et al., 2009; Bengio et al., 2012; Gangeh et al., 2015; Mairal et al., 2008; Jiang et al., 2011; Guo et al., 2012; Cai et al., 2014) and (Liu et al., 2016) for learning discriminative sparse representations have been shown to perform well across various learning tasks. A subclass of them known as discriminative dictionary learning (DDL) (Guo et al., 2012; Jiang et al., 2013; Cai et al., 2014; Shekhar et al., 2014; Xu et al., 2015; Liu et al., 2016; Bengio et al., 2012; Gangeh et al., 2015; Jiang et al., 2016) and (Vu & Monga, 2016) addresses the estimate of the dictionary in a supervised manner such that the representation w.r.t. words (vectors) from the resulting dictionary (vector set) is discriminative.

Most of the DDL methods synthesize the data sample \( k \) from class \( c \), i.e., \( x_{c,k} \in \mathbb{R}^N \) as an approximation by a linear combination \( y_{c,k} \in \mathbb{R}^M \) (referred to as a sparse data representation \( ||y_{c,k}||_0 << M \) of a few words (vectors \( d_m \)), from a dictionary (vector set) \( D \in \mathbb{R}^{N \times M} \), i.e., \( x_{c,k} = Dy_{c,k} + v_{c,k} \), \( v_{c,k} \in \mathbb{R}^N \), with \( v_{c,k} \) denoting the approximation error. It is important to highlight that with the synthesis model approach the data reconstruction is addressed.

The differences between the DDL methods (Guo et al., 2012; Jiang et al., 2013; Cai et al., 2014; Shekhar et al., 2014; Gangeh et al., 2015; Xu et al., 2015; Liu et al., 2016; Bengio et al., 2012; Jiang et al., 2016) and (Vu & Monga, 2016) are determined by the prior defined on the sparse representation and the prior defined for the relations between the sparse representations for the data samples.
from the same/different classes. The discrimination is enforced by replacing the prior with a structural constraint on the dictionary or imposing a discriminative term on the sparse representations. Additionally, some works by (Mairal et al., 2008; Guo et al., 2012) and (Jiang et al., 2015) consider even a joint estimation/learning of a dictionary, sparse representation, and classifier by using iterative alternating minimization strategy. The manuscripts by (Bengio et al., 2012; Cai et al., 2014) and (Gangeh et al., 2015) give comprehensive overview covering different approaches.

1.1 Open issues

The general open issue for DDL methods is the computational complexity w.r.t. the optimal dictionary/transform learning and the discriminative encoding, since the sparse representation in the synthesis model is a solution to an inverse problem.

An additional open issue with most of the proposed approaches (Guo et al., 2012; Jiang et al., 2013; Cai et al., 2014; Gangeh et al., 2015; Liu et al., 2016), (Bengio et al., 2012; Jiang et al., 2016; Vu & Monga, 2016) is that there is no formal notion to measure the discriminative properties. Therefore, there are no means that provide a quantitative evaluation about the quality of the representation, other then the performance of a classifier used on top of the representation.

Concerning the specifics in the discriminative constraints, (Yang et al., 2011b) proposed a synthesis model with a discriminative fidelity term and Fisher discriminant constraints, where the within-class scatter and the between-class scatter of the representation is minimized and maximized, respectively. The authors (Vu & Monga, 2016) proposed an extension considering a low-rank constraint on the dictionary. An approach by (Guo et al., 2013) used a synthesis model with constraint on the pair-wise relation between the sparse representation expressed by $\ell_2$ distance metric. The methods reported by (Yang et al., 2011b) (Vu & Monga, 2016) and (Guo et al., 2013) take into account assumption on the metric by defining the scatter and the pair-wise relations. Therefore, they constrain the space of the representation, that essentially is determined by the dictionary. However, these works do not consider whether the used metric is optimal w.r.t. the sparse representation.

The method proposed by (Liu et al., 2016) finds a dictionary under which the coefficients of a data sample from the same class $c$ have a common sparse structure while the size of the overlapping signal support of different classes, denoted as $c_1$ and $c_2$ is minimized. Furthermore, assuming $y_{c_1,k_1} \in \mathbb{R}^M$ and $y_{c_2,k_2} \in \mathbb{R}^M$ are two sparse representations for data samples $x_{c_1,k_1} \in \mathbb{R}^N$ and $x_{c_2,k_2} \in \mathbb{R}^N$, they proposed a similarity measure defined by empirical expectation on $\|y_{c_1,k_1} \odot y_{c_2,k_2}\|_0$, where $\odot$ represents the Hadamar product. Note that two transform data samples $y_{c_1,k_2}$ and $y_{c_2,k_2}$ that have small support overlap $\|y_{c_1,k_1} \odot y_{c_2,k_2}\|_0 = s \ll M$, might not necessarily be similar or dissimilar, i.e., $y_{c_1,k_1} = y_{c_2,k_2}$ and $y_{c_1,k_1} = -y_{c_2,k_2}$ with $\|y_{c_1,k_1}\|_0 = \|y_{c_2,k_2}\|_0 = s$ and $s$ small.

1.2 Approach and motivations

**Approach** Instead of addressing a synthesis model where the data reconstruction is targeted and the estimation of the discriminative representation is an inverse problem, we present a novel, alternative approach. That is we propose non-linear transform models in the learning stage and a sparsifying transform model (Rubinstein et al., 2010; Rubinstein et al., 2013; Rubinstein & Elad, 2014; Ravishankar & Bresler, 2014) for testing. The sparsifying transform model assumes that the data sample $x_{c,k}$ is approximately sparsifiable under a linear transform $A \in \mathbb{R}^{M \times N}$, i.e., $A x_{c,k} = y_{c,k} + z_{c,k}$, $z_{c,k} \in \mathbb{R}^M$, where $y_{c,k}$ is sparse $\|y_{c,k}\|_0 << M$. It also represents a generalization of the analysis model (Rubinstein et al., 2010; 2013; Ravishankar & Bresler, 2014; Rubinstein & Elad, 2014).

The proposed non-linear transform model is an extension to the sparsifying transform model that considers additional assumptions. Both of the models used in this paper address a direct problem, where the estimation of the discriminative representation represents a low complexity constrained projection problem. Additionally, since these models have no restrictions on the transform representation to be in the column space of the dictionary (transform matrix $A$), they allow more freedom in modeling and imposing constraints on the transform representation.

---

1. Note that a solution to an inverse problem has a high computational complexity if the dimensionality of the dictionary (transform matrix) or the data dimensionality is high.

2. In fact it allows modeling other non-linearity also, i.e., ReLu can be modeled as a transform representation.
Non-linear transform mapping directly put on the contributing components for similarity/dissimilarity. Consider the measuring the parametric measure with an non-parametric one the focus of the regularization is di-
properties without taking into account any additional assumptions. Second, by approximat-
tween a transform data allows more freedom in imposing regularization on the discriminative 
y 

Motivations The first motivation behind the used measure is that the support intersection be-
tween a transform data allows more freedom in imposing regularization on the discriminative properties without taking into account any additional assumptions. Second, by approximat-
ing the parametric measure with a non-parametric one the focus of the regularization is di-
rectly put on the contributing components for similarity/dissimilarity. Consider the measure 

\[ \|y_{c,k}^+ \cap y_{c2,k2}^+ \| + \|y_{c,k}^- \cap y_{c2,k2}^- \| \]

for the vectors \( y_{c,k} \) and \( y_{c2,k2} \). Note that the dissimilarity contribution

\[ -\|y_{c,k}^- \cap y_{c2,k2}^- \| - \|y_{c,k}^+ \cap y_{c2,k2}^+ \| \]

for the vectors \( y_{c,k} \) and \( y_{c2,k2} \) is zero then \( \|y_{c,k} \cap y_{c2,k2} \| = \|y_{c,k}^+ \cap y_{c2,k2}^+ \| + \|y_{c,k}^- \cap y_{c2,k2}^- \| \). Third, the expectation of this similarity measure across the dataset captures the concentration of similarity for that set. Therefore, it provides the possibility to define a formal notion of quantifying the discriminative properties.

Forth, the measure \( \|y_{c,k} \cap y_{c2,k2} \| = \|y_{c,k}^+ \cap y_{c2,k2}^+ \| + \|y_{c,k}^- \cap y_{c2,k2}^- \| \) is not ambiguous with a notion for similarity/dissimilarity between two sparse representations \( y_{c,k} \) and \( y_{c2,k2} \). This is because the support intersections for the positive and the negative components \( (y_{c,k}^+ \cap y_{c2,k2}^+ \| + \|y_{c,k}^- \cap y_{c2,k2}^- \| \) and \( (y_{c,k}^+ \cap y_{c2,k2}^+ \| + \|y_{c,k}^- \cap y_{c2,k2}^- \| \) are considered separately. In addition a measure for the strength on the support intersection, defined as \( \|y_{c,k} \cap y_{c2,k2} \| \) is also taken into account. The expected strength of the measure \( \|y_{c,k} \cap y_{c2,k2} \| \) across a data set captures the expected strength on the support intersection for that set. A schematic diagram of the transform and the main idea behind the proposed concept are shown in Figure 1, a), b) and c). On Figure 1, d), e) and f) are given illustrative examples for the support intersections between the positive and negative component of two vectors \( y_{c,k} \) and \( y_{c2,k2} \) in the transform domain.

A learning algorithm is presented using the proposed model with discriminative and minimum infor-

Figure 1: a) Data samples \( X_c, c \in \{1, 2, 3, 4\} \) from four different classes, b) given a \( k \)-th data sample \( x_{c,k} \) from class \( c \), the non-linear transform is represented as two step operation: linear mapping \( A x_{c,k} \) (step 1) followed by an element wise thresholding function \( y_{c,k} = T P_{\epsilon}(A x_{c,k}) \) (step 2), c) the transform data samples \( Y_c, c \in \{1, 2, 3, 4\} \), d) two transform representations \( y_{c1,k1} \) and \( y_{c2,k2} \), e) the resulting Hadamar products \( y_{c1,k1} \cap y_{c2,k2} \) and \( y_{c1,k1} \cap y_{c2,k2} \) on the support intersection for the similarity contribution and f) the resulting Hadamar products \( y_{c1,k1} \cap y_{c2,k2} \) and \( y_{c1,k1} \cap y_{c2,k2} \) on the support intersection for the dissimilarity contribution between \( y_{c1,k1} \) and \( y_{c2,k2} \).
1.3 Notations

A scalar variable is denoted using $x$, a vector is denoted by a bold, low caps symbols $\mathbf{x}$, a matrix by bold, upper cap symbol $\mathbf{A}$. A single element from a vector (or matrix) is denoted as $x(n)$ (or $A(m,n)$). A set is denoted by a calligraphic symbol, i.e., $\mathcal{S}$. The $\ell_p$-norm is denoted as $\|\cdot\|_p$ and the nuclear norm as $\|\|_*$. The $\odot$ symbol represents the Hadamard product. Throughout the paper it is assumed that a set of data samples $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_C\} \in \mathbb{R}^{N \times L}$, $L = \mathcal{C} K$ from $\mathcal{C}$ classes is given and that every class $c \in \mathcal{C} = \{1, 2, 3, ..., C\}$ has $K$ samples, $\mathbf{X}_c = \{\mathbf{x}_{c,1}, \mathbf{x}_{c,2}, ..., \mathbf{x}_{c,K}\} \in \mathbb{R}^{N \times K}$, $\mathbf{x}_{c,k} \in \mathbb{R}^N$, $\forall c \in \mathcal{C}, \forall k \in \mathcal{K} = \{1, 2, ..., K\}$. We denote the transform data as $\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_C\} \in \mathbb{R}^{M \times L}$, where $\mathbf{Y}_c = \{\mathbf{y}_{c,1}, \mathbf{y}_{c,2}, ..., \mathbf{y}_{c,K}\} \in \mathbb{R}^{M \times K}$ and $\mathbf{y}_{c,k} \in \mathbb{R}^M$. We denote $\mathbf{X}_{\cdot \{k \in c\}} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_{c,k}, ..., \mathbf{X}_C] \in \mathbb{R}^{N \times (L-1)}$ as the matrix that has all the columns of $\mathbf{X}$, except the column $\mathbf{x}_{c,k} \in \mathbb{R}^N$, where $\mathbf{X}_{c,k} = [\mathbf{x}_{c,1}, \mathbf{x}_{c,2}, ..., \mathbf{x}_{c,k-1}, \mathbf{x}_{c,k+1}, ..., \mathbf{x}_{c,K}] \in \mathbb{R}^{N \times (K-1)}$ is a matrix that has all the columns of block $\mathbf{X}_c$, except the column $\mathbf{x}_{c,k}$, $\forall c \in \mathcal{C}$ and $\forall k \in \mathcal{K}$. We let $\mathcal{N} = \{1, 2, ..., N\}$ and $\mathcal{M} = \{1, 2, ..., M\}$.

2 Learning Non-linear Transform with Discriminative and Minimum Information Loss Priors

The proposed approach has two operational modes: learning and test. It considers two different models. A predefined number of non-linear transforms are used in the learning mode and one sparsifying transform is used for test time.

2.1 The Parametric Non-linear Transform Modeling

Considering the learning mode, we assume that for every class $c \in \mathcal{C}$ there exist one non-linear transform defined by a set of parameters $\mathcal{P}_c = \{\mathbf{A} \in \mathbb{R}^{M \times N}, \mathbf{\tau}_c \in \mathbb{R}^M\}, \forall c \in \mathcal{C}$. The linear map $\mathbf{A}$ is shared for all of the $\mathcal{C}$ different models, but, the parameters $\mathbf{\tau}_c$ for the different classes $c$ from the corresponding non-linear transforms $\mathcal{P}_c$ are different. It is assumed that the parameters $\mathbf{\tau}_c$ are spread far apart in the transform domain. In addition, when all the non-linear transforms are applied to the corresponding class samples then the transform data samples are separable w.r.t. the different classes, in the transform domain.

As far as the non-linear transform we focus on transforms expressible by a two-steps operation, consisting of a linear mapping (step 1) followed by an element-wise non-linearity (step 2):

$$\mathbf{x}_{c,k} \xrightarrow{\text{step 1}} \mathbf{A}\mathbf{x}_{c,k} \xrightarrow{\text{step 2}} \mathbf{y}_{c,k},$$

where $\forall \mathbf{w}_{c,k} = \mathbf{A}\mathbf{x}_{c,k} \in \mathbb{R}^M$, $\mathcal{H}_{\tau_c}(\cdot)$ is the non-linear thresholding function with parameters $\mathbf{\tau}_c \in \mathbb{R}^M$, represents a non-linear thresholding function with parameters $\mathbf{\tau}_c \in \mathbb{R}^M$. We also have to mention that the thresholding is done with different thresholding parameters $\mathbf{\tau}_c(m)$ for the corresponding different transform dimensions $m \in \mathcal{M}$.

At testing time we use one, common sparsifying transform defined by a set of parameters $\mathcal{P} = \{\mathbf{A} \in \mathbb{R}^{M \times N}, \mathbf{\tau} \in \mathbb{R}^M\}$ for all data samples, with a constant thresholding parameter $\mathbf{\tau}$. The transform matrix $\mathbf{A}$ is one and same for the $\mathcal{C}$ training and the testing models.

In the following section we describe the non-linear transform models, the proposed discriminative prior and its mesure, together with the main reason behind this particular use of non-linear transform models for learning and a sparsifying transform model for testing.

2.2 The Non-linear Transform Model with a Discriminative Prior

The learning model This paper defines a compact description of the non-linear transform [1] by a non-linear transform model as follows:

$$\mathbf{A}\mathbf{x}_{c,k} = \mathbf{y}_{c,k} + \mathbf{z}_{c,k}, \quad \mathbf{y}_{c,k} = \mathcal{T}^{\mathcal{P}_c}(\mathbf{x}_{c,k}),$$

3That is the number of non-linear transforms equals the number of classes.
where $\mathcal{T}^{P_c}(\cdot) : \mathbb{R}^N \rightarrow \mathbb{R}^M$ is a parametric non-linear function that gives $y_{c,k}$, by using the set of parameters $P_c$. The term $z_{c,k} = Ax_{c,k} - y_{c,k}$ is the non-linear transform error vector that represents the deviation of $Ax_{c,k}$ from the targeted transform representation $y_{c,k} = \mathcal{T}^{P_c}(x_{c,k})$ in the transform domain. Since the transform representation $y_{c,k} = \mathcal{T}^{P_c}(x_{c,k})$ takes into account a non-linearity, $Ax_{c,k}$ is only seen as its linear approximation. Knowing something in advance about the difference between $y_{c,k} - Ax_{c,k}$ can be used in our model. However, since in advance we do not have any prior we assume that it is Gaussian like distributed. Therefore, the prior on $z_{c,k}$ is modeled as $p(z_{c,k}|y_{c,k}, A) \propto \exp(-\|Ax_{c,k} - y_{c,k}\|^2_{\beta_0})$, where $\beta_0$ is a scaling parameter. Assuming additionally that the non-linear function $\mathcal{T}^{P_c}(x_{c,k})$ gives sparse $y_{c,k}$, then we have the improper prior on $y_{c,k}$, defined as $p(y_{c,k}) \propto \exp(-\|y_{c,k}\|^2_{\beta_1})$ where $\beta_1$ is a scaling parameter. This paper models the joint probability $p(\tau_1, \tau_2, ..., \tau_C; y_{c,k})$ as

$$
p(\tau_1, \tau_2, ..., \tau_C; y_{c,k}) \propto \exp\left(-\frac{1}{\beta_2} \min_{1 \leq c \leq C} D(y_{c,k}; \tau_c)\right) p(y_{c,k}),$$

and assumes that $(A_{iid}) : p(\tau_1, \tau_2, ..., \tau_C) = \prod_c p(\tau_c)$, where $\beta_2$ is a scaling parameter. By the assumption in subsection 2.1, $\tau_1$ are spread far apart in the transform domain, therefore, a minimum over $D(y_{c,k}; \tau_1)$ ensures that $y_{c,k}$ in the transform domain will be located to the closest $\tau_1$ w.r.t. to the measure $D(y_{c,k}; \tau_1)$. Moreover, given the class label $c$ using $(A_{iid})$ we model the discriminative prior as $p(\tau_c|y_{c,k}) \propto \exp\left(-\frac{D(y_{c,k}; \tau_c)}{\beta_2}\right)$ by using a parametric measure $D(y_{c,k}; \tau_c)$ with parameter $\tau_c$. Assuming that $D(y_{c,k}; \tau_c)$ is determined by a relation on the support intersection between $y_{c,k}$ and $\tau_c$ we propose the following definition of measure:

$$D(y_{c,k}; \tau_c) = \|y_{c,k}^+ \cap \tau_c^+\|_1 + \|y_{c,k}^- \cap \tau_c^-\|_1 + \|y_{c,k} \cap \tau_c\|^2_2,$$

where $\|y_{c,k}^+ \cap \tau_c^+\|_1 + \|y_{c,k}^- \cap \tau_c^-\|_1$ measures the similarity contribution on the support intersection using the positive and the negative components, $y_{c,k}^+ \cap \tau_c^+$ and $y_{c,k}^- \cap \tau_c^-$, respectively, of $y_{c,k}$ and $\tau_c$ and $\|y_{c,k} \cap \tau_c\|^2_2$ measures the strength of the support intersection between $y_{c,k}$ and $\tau_c$. The true $p(\tau_c)$ and $\tau_c$ are not known and instead of estimating them explicitly, an approximation to $D(y_{c,k}; \tau_c)$ is considered based only on the concentrations of the similarity on the support intersection and the expected strength of the support intersection for the transform data.

**Non-parametric approximation** We propose an approximation by sum of two expectations. The first one is the expected similarity on the support intersection for the positive and negative component between all $y_{c,k}$ and the corresponding sets of the transform representations $Y_{\setminus c}$ that come from all classes $c_1$ different from $c$, i.e., $c \neq c_1$. The second is the expected strength on the support intersection between $y_{c,k}$ and the set of transform representations $Y_{\setminus c}$ that come from all classes $c_1$ different from $c$, $c \neq c_1$. We define the approximation as:

$$\sum_{c,k} D(y_{c,k}; \tau_c) \sim D_{\ell_1}^P(X) + S_{\ell_2}^P(X) \text{ where}
$$

$$D_{\ell_1}^P(X) = \sum_{c=1}^C \sum_{k=1}^K \sum_{c_1 \in \{1,2,\ldots,C\} \setminus c} \sum_{k_1=1}^K (\|y_{c,k}^+ \cap y_{c_1,k_1}^+\|_1 + \|y_{c,k}^- \cap y_{c_1,k_1}^-\|_1),$$

$$S_{\ell_2}^P(X) = \sum_{c=1}^C \sum_{k=1}^K \sum_{c_1 \in \{1,2,\ldots,C\} \setminus c} \sum_{k_1=1}^K \|y_{c,k} \cap y_{c_1,k_1}\|^2_2,$$

we highlight that the transform representations $y_{c,k}$ used in the approximation $D_{\ell_1}^P(X)$ and $S_{\ell_2}^P(X)$ are the result of applying the sparsifying transform with parameter set $P$ to the data samples $x_{c,k}$. The $m$-th element $y_{c,k}^+(m)$ of $y_{c,k}^+$ is defined as $y_{c,k}^+(m) = \max\{y_{c,k}(m), 0\}$ and similarly, $y_{c,k}^-(m) = \max\{-y_{c,k}(m), 0\}$, $\forall m \in \mathcal{M}$. We also define the expected similarity using the positive and negative components of all $y_{c,k}$ across the transform representations $Y_c$ that come from the same classes $c$ as $D_{\ell_1,c}^P(X) = \sum_{c} \sum_{k=1}^K \sum_{k_1 \in \{1,2,\ldots,K\} \setminus c} (\|y_{c,k}^+ \cap y_{c_1,k_1}^+\|_1 + \|y_{c,k}^- \cap y_{c_1,k_1}^-\|_1)$. If the measure $D_{\ell_1}^P(X)$ is not used then the approximation [5] is most similar to the one proposed in Liu et al. (2016).
By (5) a link is established between the non-linear transform models and the sparsifying transform model, or more generally, assuming $A$ is known, between a parametric and a non-parametric modeling view. The terms

$$
\frac{2}{[(C-1)K]|c|C(K)} \sum_{c \in \{1, 2, \ldots, C\} \setminus \{c\}} \sum_{k=1}^{K} y_{c,1,k}^{+} + \frac{2}{[(C-1)K]|c|C(K)} \sum_{c \in \{1, 2, \ldots, C\} \setminus \{c\}} \sum_{k=1}^{K} y_{c,1,k}^{-}
$$

are seen as finite sample estimates of the positive $\tau_c^+$, negative $\tau_c^-$ components and the the Hadamard square $\tau_c \cdot \tau_c$ for the unknown variable $\tau_c$, $\forall c \in \mathcal{C}$.

Note that the Fisher discriminate constraint (Yang et al., 2011b), the pairwise constraint (Guo et al., 2013) and the support intersection constraint (Liu et al., 2016) are all approximations of a discriminative prior. However, they all have a specific assumptions on the distribution of the data representation in the transform domain. The advantage of using (5) is that the approximation is without any prior on the probability distributions $p(\tau_c)$ and without any explicit assumption about the distance metric/measure, or the space manifold in the transform domain.

Given a training set $X$, the learning perspective w.r.t. the discriminative properties is to estimate a non-linear transform models $P_c = \{A \in \mathbb{R}^{M \times N} \text{, } \tau_c \in \mathbb{R}^M\}$ that minimize the empirical expectation $\mathbb{E}[\min_{1 \leq c \leq C} D(y_{c,k}; \tau_c)] \approx \frac{1}{CK} \sum_{c,k} \min_{1 \leq c \leq C} D(y_{c,k}; \tau_c)$ where $\tau_c$ are known, then $\sum_{c,k} \min_{1 \leq c \leq C} D(y_{c,k}; \tau_c)$ equals to $\sum_{c,k} D(y_{c,k}; \tau_c)$ and exactly matches the empirical expectation $D_{l_1}(X) + S_{l_2}(X)$. Moreover, if the corresponding class labels for the training set $X$ are given then $\sum_{c,k} \min_{1 \leq c \leq C} D(y_{c,k}; \tau_c)$ is zero, i.e., $D(y_{c,k}; \tau_c) = 0$, then the discriminative prior is non-informative, in a sense that it has no influence in the models. Only then do the non-linear transforms reduce to the sparsifying transform model, since (6) reduce to:

$$
\min_{y_{c,k}} \|Ax_{c,k} - y_{c,k}\|^2_2 + \lambda_1 \|y_{c,k}\|_1, \tag{6}
$$

The testing model Assume that the transform matrix $A$ and the parameters $\tau_1, \tau_2, \ldots, \tau_C$ are known, then given any data sample $x_{c,k}$ the transform representation $y_{c,k}$ is estimated as:

$$
\min_{y_{c,k}} \|Ax_{c,k} - y_{c,k}\|^2_2 + \lambda_0 (\min_{1 \leq c \leq C} D(y_{c,k}; \tau_c)) + \lambda_1 \|y_{c,k}\|_1, \tag{6}
$$

which is equal to $\min_{1 \leq c \leq C} (\min_{y_{c,k}} \|Ax_{c,k} - y_{c,k}\|^2_2 + \lambda_0 D(y_{c,k}; \tau_c) + \lambda_1 \|y_{c,k}\|_1)$. Furthermore, if we assume that the measures $D(y_{c,k}; \tau_c)$ are zero, i.e., $D(y_{c,k}; \tau_c) = 0$, then the discriminative prior is non-informative, in a sense that it has no influence in the models. Only then do the non-linear transforms reduce to the sparsifying transform model, since (6) reduce to:

$$
\min_{y_{c,k}} \|Ax_{c,k} - y_{c,k}\|^2_2 + \lambda_1 \|y_{c,k}\|_1. \tag{7}
$$

Considering the testing stage, we note that the result (5) sheds light on another view. Namely, the sparsifying transform model $P = \{A \in \mathbb{R}^{M \times N}, \lambda_1 \in \mathbb{R}^M\}$ is also seen as an approximation to the models represented by a set of parameters $P_c = \{A \in \mathbb{R}^{M \times N}, \tau_c \in \mathbb{R}^M\}$ with expected loss in the discriminative properties of the transform representations expressed by the similarity concentration measure $D_{l_1}(X) + S_{l_2}(X)$. At the same time this measure can also be considered as an empirical risk (Vapnik, 1995) w.r.t. the discriminative properties, related to the generalization capabilities (Vapnik, 1995) and (Mark, 2010) of the sparsifying transform model. Note that the same model is also the simplest that approximates the non-linear transform models used in the learning stage.

### 2.3 The learning algorithm

In summary, the used priors are:

$$
p(x_{c,k}|y_{c,k}, A) \propto \exp(-\frac{\|Ax_{c,k} - y_{c,k}\|^2_2}{\beta_0})
$$

$$
p(\tau_c|y_{c,k}, p(y_{c,k}) \propto \exp(-\frac{D(y_{c,k}; \tau_c)}{\beta_2}) \exp(-\frac{\|y_{c,k}\|_1}{\beta_1}). \tag{8}
$$

---

4 Since $\sum_{c \in \{1, 2, \ldots, C\} \setminus \{c\}} \sum_{k} |y_{c,k}^+ \ominus y_{c,1,k}^+| + |y_{c,k}^- \ominus y_{c,1,k}^-| = \| (\sum_{c \in \{1, 2, \ldots, C\} \setminus \{c\}} \sum_{k} y_{c,k}^+ \|_1 + \| (\sum_{c \in \{1, 2, \ldots, C\} \setminus \{c\}} \sum_{k} y_{c,k}^- \|_1 \cdot \tau_c \sim \frac{2}{[(C-1)K]|c|C(K)} \sum_{c \in \{1, 2, \ldots, C\} \setminus \{c\}} \sum_{k} y_{c,1,k}^+ + \frac{2}{[(C-1)K]|c|C(K)} \sum_{c \in \{1, 2, \ldots, C\} \setminus \{c\}} \sum_{k} y_{c,1,k}^-

5 Since $\sum_{c \in \{1, 2, \ldots, C\} \setminus \{c\}} \sum_{k} |y_{c,k} \ominus y_{c,1,k}|^2 = (\sum_{c \in \{1, 2, \ldots, C\} \setminus \{c\}} \sum_{k} y_{c,1,k}^+ \ominus y_{c,1,k}^-)^2$

6 Note that if the labels are not given then the unsupervised case can also be addressed by using a likelihood measure between a sample and the rest of the available samples, with possibility to be defined in the original or in the transform domain.
Additionally, we have a prior on A that penalizes the information loss in order to avoid trivially unwanted matrices A, i.e., matrices that have repeated or zero rows. The prior is defined as:

\[
p(A) \propto \exp(-\Omega(A)) = \exp \left( -\frac{1}{\beta_3} \|A\|_F^2 + \frac{1}{\beta_4} \|AA^T - I\|_F^2 - \frac{1}{\beta_5} \log |\det A^T A| \right),
\]

where the \(\|A\|_F\) penalty helps regularize the scale ambiguity, the log |\det (A^T A)| and \(\|A\|_F^2\) are functions of the singular values of A and together help regularize the conditioning of A. Assuming that the expected coherence \(\mu^2(A)\) between the rows \(a_m\) of A (i.e., \(A^T = [a_1, a_2, \ldots, a_M]\)) is defined as \(\mu^2(A) = \frac{2}{M(M-1)} \sum_{m_1 \neq m_2} |a_m a_{m_2}|^2\), \(\forall m_1, m_2 \in \{1, 2, \ldots, M\}\). Then \(\|AA^T - I\|_F^2\) measures the expected coherence \(\mu^2(A)\) and the \(\ell_2\) norm for the rows of A.

Note that the joint probability can be expressed as:

\[
p(x_{c,k}, y_{c,k}, \tau_c, A) = p(x_{c,k}, y_{c,k}, \tau_c | A)p(A),
\]

where \(p(x_{c,k}, y_{c,k}, \tau_c | A) = p(x_{c,k} | y_{c,k}, A)p(\tau_c | y_{c,k}, A)p(y_{c,k} | A)\) since \(p(x_{c,k} | y_{c,k}, \tau_c, A) = p(x_{c,k} | y_{c,k}, A)\). Given the available training data set \(X\), maximizing \(p(x_{c,k}, y_{c,k}, \tau_c, A)\) over \(Y\) and \(A\) is same as minimizing the following problem:

\[
\min_{Y, A} \|AX - Y\|_2^2 + \sum_{c,k}\lambda_0 D(y_{c,k}; \tau_c) + \lambda_1 \|y_{c,k}\|_1 + \Omega(A),
\]

where \(\lambda_0, \lambda_1\) are inversely proportional to the scaling parameters \(\beta_2, \beta_1\). Note that the solution to (11) is not equivalent to the maximum a priori (MAP) solution, which is difficult to compute, as it involves integrating over the vectors \(y_{c,k}\). Considering the optimization perspective, the problem is not convex in the variables \((Y, A)\). The proposed solution here is obtained by iteratively, marginally maximizing the probability \(p(x_{c,k}, y_{c,k}, \tau_c, A)\) over \(Y\) and \(A\) which is equivalent to maximizing the conditional densities \(p(y_{c,k} | x_{c,k}, \tau_c, A)\) and \(p(A | x_{c,k}, y_{c,k}, \tau_c)\), respectively. Meaning that at one iterating step one of the variables \(Y\) or \(A\) is fixed and w.r.t. the other the problem (11) is minimized. The following describes the iterating steps that consist of linear map estimation (maximizing \(p(A | x_{c,k}, y_{c,k}, \tau_c))\) and discriminative encoding (maximizing \(p(y_{c,k} | x_{c,k}, \tau_c, A))\).

**Linear map estimation:** Given the available data samples \(X\) and the corresponding transform representations \(Y\) the linear map \(A\) estimation problem reduces to:

\[
\min_A \|AX - Y\|_2^2 + \frac{\lambda_2}{2} \|A\|_F^2 + \frac{\lambda_3}{2} \|AA^T - I\|_F^2 - \lambda_4 \log |\det A^T A|,
\]

where \(\lambda_2, \lambda_3, \lambda_4\) are inversely proportional to the scaling parameters \(\beta_3, \beta_4, \beta_5\) and we use the \(\epsilon\)-close closed form solution estimated as follows:

**Proposition 1** (\(\epsilon\)-close closed form solution): Given \(Y \in \mathbb{R}^{M \times CK}, \forall X \in \mathbb{R}^{N \times CK}\) and \(M \geq N\), \(\forall \lambda_2 \geq 0, \lambda_3 \geq 0, \lambda_4 \geq 0\) let the eigen value decomposition \(U_X \Sigma_X V_X^T\) of \(XX^T + \lambda_2 I\) and the singular value decomposition \(U_{X|XY} \Sigma_{X|XY} V_{X|XY}^T\) of \(U_X X_{XY}^T\) exist, then if and only if \(\sigma_X(n) > 0, \forall n \in N = \{1, 2, 3, \ldots, N\}\), (12) has an approximate solution as:

\[
A = U_{X|XY} V_{X|XY}^T \Sigma_A^{-1} U_X^T,
\]

where \(\Sigma_A\) is diagonal matrix, \(\Sigma_A(n,n) = \sigma_A(n) \geq 0\), and \(\sigma_A(n)\) are solutions to \(\min_{\sigma_A(n)} \frac{\lambda_3 \lambda_2}{\sigma_A(n)} \sigma_A^2(n) + \frac{\lambda_3 - 2 \lambda_2}{\sigma_A(n)} \sigma_A(n) - \frac{\lambda_4}{\sigma_A(n)}\), with \(\sigma_T(n) = T(n, n)\).

**Discriminative encoding:** Given the available data samples \(X\) and the current estimate of the transform \(A\) the discriminative representation estimation problem is formulated as \((P_{DR})\):

\[
\min_{Y_{c,k}} \|AX_{c,k} - Y_{c,k}\|_2^2 + \sum_{c,k} \lambda_0 D(y_{c,k}; \tau_c) + \lambda_1 \|y_{c,k}\|_1.
\]

Note that (11) and consequently \((P_{DR})\) is a well defined problem only if the model variables \(\tau_c\), or the probability distributions \(p(\tau_c)\) are known in advance. Nevertheless, we show that the approximation (13) leads to an efficient solution of \((P_{DR})\). Assuming that \(Y_{c,k}\) is given, then for any sample \(k \in C\) from any class \(c \in C\) the approximation^^^leads to a constrained projection problem:

\[
\min_{Y_{c,k}} \|AX_{c,k} - Y_{c,k}\|_2^2 + \lambda_0 D(y_{c,k}; \tau_c) + \lambda_1 \|y_{c,k}\|_1.
\]
and has a closed form solution as:

\[ y_{c,k} = \text{sign}(A_{x_c,k}) \odot \max \left( (|Ax_{c,k}| - \lambda_0g + \lambda_11,0) \odot (1 + 2|g_0s_c) \right), \]

where \( \odot \) is a Hadamard division, \( g = \text{sign}(\max(|Ax_{c,k},0|)) \odot d^+ + \text{sign}(\max(-Ax_{c,k},0)) \odot d^- \),

\[ d_c^+ = \sum_{c_1 \neq c} y_{c_1,k}^+ \quad d_c^- = \sum_{c_1 \neq c} y_{c_1,k}^- \quad \text{and} \quad s_c = \sum_{c_1 \neq c} \sum_{k_1} y_{c_1,k_1} \odot y_{c_1,k_1} \]

(15)

Theorem 1: The bound on the discrimination power is given by the following result.

\[ \log(\lambda_{\min}(A^T A)) + \log \left( \frac{\partial D_{B_1,c}(X)}{\partial A_{u,v} | A_{u,v}=1} \right) \leq \mathcal{T}^I \leq \log \left( D_{B_1,c}^{B_1}(AX) \right) - \log \epsilon. \]

(17)

The proof is given in Appendix D. 

\[ \begin{align*}
\mathcal{T}^I &= \log(D_{B_1,c}^{B_1}(Y)) - \log(D_{B_1,c}^{B_1}(Y) + \epsilon) - \log(D_{B_1,c}(X)) - \log(D_{B_1,c}(X) + \epsilon), \\
\mathcal{T}^p &= \log \left( D_{B_1,c}^{B_1}(AX) \right) - \log \left( D_{B_1,c}^{B_1}(Y) \right) - \log \epsilon.
\end{align*} \]

(16)

We note that at convergence (which we do not prove here) we can only claim that a joint local maximum in \((Y, A)\) of \(p(x_c,y,c,\tau_c,A)\) has been reached, even if, in this case, each optimization step achieves the (marginal) \(\epsilon\)-close and the global optimal solution, respectively.

2.4 A MEASURE FOR THE DISCRIMINATIVE PROPERTIES AND ITS BOUNDS

This paper proposes a notion for the discriminative properties of a data set under a non-linear transform named as discrimination power, based on a measure for the relations between the concentrations \(D_{B_1,c}^{B_1}(X)\) and \(D_{B_1,c}(X)\).

Proposition 2: The discrimination power for any dataset \(Y \in \mathbb{R}^{M \times CK}\) under a non-linear transform with parameter set \(P\) is defined as:

\[ \mathcal{T}^I = \log(D_{B_1,c}^{B_1}(Y)) - \log(D_{B_1,c}^{B_1}(Y) + \epsilon) - \log(D_{B_1,c}(X)) - \log(D_{B_1,c}(X) + \epsilon), \]

where \(B^M = \{A_1 \in \mathbb{R}^{M \times M}, 0 \in \mathbb{R}^M\}\), \(A_1 = I\) and \(\epsilon > 0\) is a small positive constant.

Remark 1: The advantage of this measure is that it logarithmically signifies the difference between \(D_{B_1,c}^{B_1}(X)\) and \(D_{B_1,c}(X)\).

The definition about the discrimination power of the original data set \(X\), but, now under a model with a parameter set \(B^N = \{A_o \in \mathbb{R}^{N \times N}, 0 \in \mathbb{R}^N\}\) where \(A_o = I\) is equivalent to the one defined for \(\mathcal{T}^I\). We denote it as \(\mathcal{T}^p\). The bound on the discrimination power is given by the following result.

Theorem 1: The discrimination power for any data set \(X \in \mathbb{R}^{N \times CK}\) under any non-linear transform with parameter set \(P\) is bounded as:

\[ \log(\lambda_{\min}(A^T A)) + \log \left( \frac{\partial D_{B_1,c}(X)}{\partial A_{u,v} | A_{u,v}=1} \right) \leq \mathcal{T}^I \leq \log \left( D_{B_1,c}^{B_1}(AX) \right) - \log \epsilon. \]

At first the resulting bounds might look counterintuitive since the loss of information seems to increase the discrimination power. This fact is true, however, up to a certain limit. Therefore, it is important to distinguish two main conclusions. First, for any model with a set of parameters \(P\) for which there is no loss of information, that is, no thresholding, the only condition for the increase in the discrimination power is \(D_{B_1,c}^{B_1}(X) \geq D_{B_1,c}^{B_1}(AX)\) and \(D_{B_1,c}(X) \leq D_{B_1,c}(AX)\). Second, in the rest of the cases for which \(D_{B_1,c}^{B_1}(X) \geq D_{B_1,c}^{B_1}(Y)\) and \(D_{B_1,c}(X) \leq D_{B_1,c}(Y)\) holds true it will be possible to increase the discrimination power. Moreover, there is a trade-off between the increase in discrimination power as a result of the loss of information as consequence of the non-linear thresholding operation.

\[ \text{Assume we have a discriminative prior defined as } p(\tau_c|y_{c,k}) \propto \exp(-\frac{||\tau_c^+ \odot y_{c,k}^+||_1 + ||\tau_c^- \odot y_{c,k}^-||_1}{\beta_2}) \text{ and } p(v_c|y_{c,k}) \propto \exp(-\frac{||v_c^+ \odot y_{c,k}^+||_1 + ||v_c^- \odot y_{c,k}^-||_1}{\beta_2}), \text{ where } \tau_c \text{ and } v_c \text{ are unknown parameters. Then the difference } D_{B_1,c}(X) - D_{B_1,c}(X) \text{ between } D_{B_1,c}(X) \text{ and } D_{B_1,c}(X) \text{ actually represents a finite sample approximation to a discriminative density since it approximates the density } \log \left( \frac{p(\tau_c|y_{c,k})}{p(v_c|y_{c,k})} \right) \text{, i.e., } D_{B_1,c}(X) - D_{B_1,c}(X) \sim -\log \left( \frac{p(\tau_c|y_{c,k})}{p(v_c|y_{c,k})} \right) \text{ and } -\log \left( \frac{p(\tau_c|y_{c,k})}{p(v_c|y_{c,k})} \right) = D(\tau_c, y_{c,k}) - D(v_c, y_{c,k}) = ||\tau_c^+ \odot y_{c,k}^+||_1 + ||\tau_c^- \odot y_{c,k}^-||_1 - (||v_c^+ \odot y_{c,k}^+||_1 + ||v_c^- \odot y_{c,k}^-||_1). \]

\[ \text{In Appendix C we also provide a sensitivity analysis that complements our result since it is related to the notion of the discriminative power of the representation. The result in Appendix C also gives an information-theoretic interpretation and information-geometric perspective about the model and the similarity concentration measure without the need of strict conditions for regularity, i.e., smoothness of the manifolds.} \]
The numerical experiments are summarized in two different parts. In the first series of the experiments the properties of the learned map $A$ are investigated. We evaluate the computational efficiency, as run time $t_e[min]$ in minutes of the proposed algorithm for 28 iterations at the transform domain dimensionality $M = 19000$.

Table 1: The conditioning number $C_n(A) = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}$ and the expected mutual coherence $\mu(A)$ for the learned transform $A$. The execution time $t_e[min]$ in minutes of the proposed algorithm for 28 iterations at the transform domain dimensionality $M = 19000$.

<table>
<thead>
<tr>
<th>$C_n(A)$</th>
<th>$\mu(A)$</th>
<th>$t_e[min]$</th>
<th>$T^O$</th>
<th>$T^{RT}$</th>
<th>$T^{ST^*}$</th>
<th>$T^{NT}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D1$</td>
<td>2.21</td>
<td>0.03</td>
<td>5.10</td>
<td>0.03</td>
<td>0.18</td>
<td>0.68</td>
</tr>
<tr>
<td>$D2$</td>
<td>1.80</td>
<td>0.02</td>
<td>5.45</td>
<td>0.02</td>
<td>0.10</td>
<td>1.30</td>
</tr>
<tr>
<td>$D3$</td>
<td>2.12</td>
<td>0.02</td>
<td>6.55</td>
<td>0.00</td>
<td>0.01</td>
<td>0.71</td>
</tr>
<tr>
<td>$D4$</td>
<td>0.08</td>
<td>0.02</td>
<td>8.92</td>
<td>0.08</td>
<td>0.61</td>
<td>0.89</td>
</tr>
<tr>
<td>$D5$</td>
<td>6.01</td>
<td>0.01</td>
<td>12.8</td>
<td>0.01</td>
<td>0.16</td>
<td>1.02</td>
</tr>
<tr>
<td>$D6$</td>
<td>33.1</td>
<td>0.02</td>
<td>30.1</td>
<td>0.06</td>
<td>0.53</td>
<td>1.36</td>
</tr>
<tr>
<td>$D7$</td>
<td>1.60</td>
<td>0.02</td>
<td>5.00</td>
<td>0.13</td>
<td>0.63</td>
<td>1.06</td>
</tr>
</tbody>
</table>

3 Numerical experiments

The numerical experiments are summarized in two different parts. In the first series of the experiments the properties of the learned map $A$ for the proposed algorithm are investigated. We evaluate the computational efficiency, as run time $t_e[min]$, the conditioning number $C_n(A) = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}$, the expected mutual coherence $\mu(A)$ and the discrimination power across several databases for a learned non-linear transform having different dimensionality. A comparison between the discrimination power under different transforms is presented. The discrimination power is estimated in the original domain, after transform by a random matrix (having Gaussian random samples as entries and transform dimension of $M = 19000$) and after a learned non-linear transform having transform dimension $M = 19000$ without and with discriminative prior, denoted as $T^O$, $T^{RT}$, $T^{ST^*}$ and $T^{NT}$, respectively. The second part evaluates a comparison of the discrimination power between the proposed algorithm and different DDL methods (Ramirez et al., 2010), (Yang et al., 2011a), (Vu et al., 2015) and (Vu & Monga, 2016). This comparison considers a setup where the used data sets are divided into a training and test set. Moreover, the learning is performed on the training set and the evaluation is performed on the test set. In the same series of experiments the recognition accuracy for the two data sets is also computed and compared.

Data sets and algorithms set up: The used data sets are Extended YALE B ($D1$) (Georghiades et al., 2001), AR ($D2$) (Martinez & Benavente, 1998), Norb ($D3$) (LeCun et al., 2004), Coil-20 ($D4$)
The algorithm is initialized with a random matrix having i.i.d. Gaussian (zero mean, variance one) entries and is terminated after the 15th iteration. The results are obtained as the average of 3 runs. An implementation presented in [Vu & Monga, 2016] was used to learn the dictionaries and estimate the singular value decomposition for a large matrix has high computational complexity. However, one simple on-line variant for the update of $\mathbf{A}$ w.r.t. a subset of the available training set has the form $\mathbf{A}^{t+1} = \mathbf{A}^t - \rho (\mathbf{A}^t - \hat{\mathbf{A}}^t)$ with $\rho$ a predefined step size. In the numerical experiments we use the on-line variant of the algorithm (the convergence analysis for this variant of the algorithm is left for future work) were we used a baches of sizes equal to 10%–12% of the total amount of the available training data. The parameters $\lambda_0$ and $\lambda_1$ are set such that the resulting non-linear transform representation has a very small number of non-zeros w.r.t. the transform dimension. In the experiments this number is set to be 15. The rest of the parameters are set as $\{\lambda_2, \lambda_3, \lambda_4\} = \{1000000, 1000000, 1000000\}$. The rest of the parameters are set as $\{\lambda_0, \lambda_1\}$ and the expected mutual coherence $\mu(\mathbf{A})$ for the learned linear transform $\mathbf{A}$ at different dimensionality $M \in \mathbb{Q}$.

Figure 3: The conditioning number $C_n(\mathbf{A}) = \frac{\lambda_{max}}{\lambda_{min}}$ and the expected mutual coherence $\mu(\mathbf{A})$ for the non-linear transform at different dimensionality $M \in \mathbb{Q}$.

Figure 4: The similarity concentrations $C_1 = D_{\mathbf{A}^t, \mathbf{Y}}^{B_{\mathbf{H}}} \mathbf{Y}$ and $C_2 = D_{\mathbf{A}^t, \mathbf{Y}}^{B_{\mathbf{H}}} \mathbf{Y}$, their ratio $C_1/C_2$ and the discrimination power $\log(C_1/C_2) = I^t$ on a subset of the transform data using learned non-linear transform at different dimensionality $M \in \mathbb{Q}$.

[1] Nene et al. (1996), Clatech101 (D5) [LeCun et al. (2008), UKB (D6) [Nistér & Stewénius (2006) and MNIST (D7) (Lecun & Cortes). All the images from the respective datasets were downsampled to resolutions $21 \times 21, 32 \times 28, 24 \times 24, 20 \times 25, 21 \times 21, 20 \times 25, 28 \times 28$, respectively, and are normalized to unit variance. Considering the used implementation of the algorithm we note that the singular value decomposition for a large matrix has high computational complexity. However, $\mathbf{A} - \hat{\mathbf{A}}$, where $\hat{\mathbf{A}}$ is estimated as a solution in the transform update step, can be considered as an proximal operator (Parikh & Boyd, 2014) for the gradient of the objective (12). Additionally, instead of using all of the available data samples $\mathbf{X}$, a subset of them might be used. Therefore, one simple on-line variant for the update of $\mathbf{A}$ w.r.t. a subset of the available training set has the form $\mathbf{A}^{t+1} = \mathbf{A}^t - \rho (\mathbf{A}^t - \hat{\mathbf{A}}^t)$ with $\rho$ a predefined step size. In the numerical experiments we use the on-line variant of the algorithm (the convergence analysis for this variant of the algorithm is left for future work) were we used a baches of sizes equal to 10%–12% of the total amount of the available training data. The parameters $\lambda_0$ and $\lambda_1$ are set such that the resulting non-linear transform representation has a very small number of non-zeros w.r.t. the transform dimension. In the experiments this number is set to be 15. The rest of the parameters are set as $\{\lambda_2, \lambda_3, \lambda_4\} = \{1000000, 1000000, 1000000\}$. The algorithm is initialized with a random matrix having i.i.d. Gaussian (zero mean, variance one) entries and is terminated after the 28th iteration. The results are obtained as the average of 3 runs. An implementation presented in [Vu & Monga, 2016] was used to learn the dictionaries and estimate the
Table 2: a) The discrimination power for the methods DLSI (Ramirez et al., 2010), FDDL (Yang et al., 2011a), COPAR (Vu et al., 2015) and LRSDL (Vu & Monga, 2016) and the proposed method NT; b) and c) The recognition results on the Extended Yale B and MNIST database.

<table>
<thead>
<tr>
<th>Method</th>
<th>D1</th>
<th>D7</th>
<th>D1 Acc. [%]</th>
<th>D7 Acc. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLSI</td>
<td>0.71</td>
<td>0.67</td>
<td>96.5</td>
<td>98.74</td>
</tr>
<tr>
<td>FDDL</td>
<td>0.87</td>
<td>0.63</td>
<td>97.5</td>
<td>96.31</td>
</tr>
<tr>
<td>COPAR</td>
<td>0.57</td>
<td>0.54</td>
<td>98.3</td>
<td>96.41</td>
</tr>
<tr>
<td>LRSDL</td>
<td>0.42</td>
<td>0.40</td>
<td>98.7</td>
<td></td>
</tr>
<tr>
<td>NT</td>
<td><strong>0.98</strong></td>
<td><strong>0.81</strong></td>
<td><strong>99.7</strong></td>
<td><strong>99.02</strong></td>
</tr>
</tbody>
</table>

Figure 5: The recognition results and the discrimination power on the Extended Yale B and MNIST databases, respectively, using a non-linear transform with different dimensionality $M$ and linear SVM classifier on top of the transform representation.

Figure 6: The expected loss $E[\|z_{c,k}\|_2^2] = E[\|Ax_{c,k} - y_{c,k}\|_M^2]$ and the discrimination power on the Extended Yale B and MNIST databases, respectively, on the transform representation $Y$, obtained by using a non-linear transform $T^P$ at different dimensionality $M$.

Linear map properties, the similarity concentrations and the discrimination power The conditioning number and the expected coherence for the learned transforms are shown on Table 1. The learned transforms for all the databases have good conditioning numbers and low expected coherence. The running time $t_e$, measured in minutes, and the number of used dimensions, denoted as $M$ are also shown in Table 1. The learned transforms for all the data sets have relatively low execution time, regardless of the very high transform dimension $M = 19000$. The discrimination power is significantly increased in the transform domain $T^NT$ compared to the one in the original domain $T^D$ and is higher than $T^ST$ and $T^RT$. The evolution of the similarity concentrations $C1 = D_{I_1,c}^{BM}(Y)$ and $C2 = D_{I_1}^{BM}(Y)$, their ratio $C1/C2$ and the discrimination power $\log(C1/C2) = I^I$ for subsets of the used databases after applying a non-linear transform with transform dimension $M = 19000$ is shown in Figure 2. It is important to note that the similarity concentrations $C1 = D_{I_1,c}^{BM}(Y)$ and
$C^2 = D^{B_M}_t(Y)$ are decreasing, meaning that there is a loss of information. However how this loss effects the resulting similarity concentration is crucial for the discrimination properties. As shown in Figure 2 the slope of decrease for $C^2$ is stronger. Therefore, the discrimination power increases per iteration. For the Coil-20 (D4) database there is a fluctuation. This is explained by the fact that during learning we used a small number of data samples from the same database and that in the data there is high variability. The conditioning number and the expected coherence for the learned transforms for all the databases at different transform dimensions $M \in Q = \{100, 1150, 2200, 3250, 4300, 5350, 6400, 7450, 8500, 9500, 10600, 11650, 12700, 13750, 14800, 15850, 16900, 17950, 19000 \}$ are shown in Figure 3. We see that the value of both the conditioning number and the coherence is reducing and it converges to common values. This confirms the effectiveness of the conditioning and the coherence constraints. The similarity concentrations $C_1 = D^{B_M}_t(Y)$ and $C^2 = D^{B_M}_t(Y)$, their ratio $C_1/C^2$ and the discrimination power log($C_1/C^2$) = $T^2$ for a subsets of the used databases after applying a non-linear transform having transform dimensions $M \in Q$ is shown in Figure 4. We can see similar behavior as previous, that is, $C_1$ and $C^2$ are decreasing, but, the slope of decrease for $C^2$ is stronger. Therefore, the discrimination power increases as the transform dimension increases.

NT vs DDL discrimination power and recognition performance The proposed method is compared with DLSI (Ramirez et al., 2010), FDDL (Yang et al., 2011a), COPAR (Vu et al., 2015) and LRSSDL (Vu & Monga, 2016). Half of the data samples from the data set Extended YALE B, sampled at random are used for learning and the remaining other half are used for evaluation. Considering the MNIST database the training set is used for learning and the test set is used for both evaluating the discrimination power and the recognition accuracy. We compute both the discrimination power and the recognition accuracy on a subsets from the test sets. The dictionary size (transform dimension $M$) is set to be equal to $\{150, 75, 1515, 3825, 570, 150, 300\}$ for the used databases, respectively, in all of the comparing algorithms. The discrimination power of the comparing methods is denoted as $T^{DLSSI}$, $T^{FDDL}$, $T^{COPAR}$ and $T^{LRSSDL}$. The recognition results for the methods DLSI, FDDL, COPAR and LRSSDL on the data sets Extended YALE B and MNIST were not computed here, rather we use the best reported result form the respective papers (Ramirez et al., 2010), (Yang et al., 2011a), (Vu et al., 2015) and (Vu & Monga, 2016). Considering the proposed algorithm the non-linear transform was learned for the transform dimensions $M = \{100, 500, 1500, 4000\}$ and $M = \{1000, 4000, 6000, 12000\}$, respectively, for the used data sets. After the transform was learned, the transform data samples were computed for the respective training and test sets. Then, the transform training data samples were used as features to learn a linear SVM classifier in one-against-all regime. The results are shown in Table 2(a), b) and c). The discrimination power of the proposed non-linear transform is higher that the discrimination power of the comparing methods. The recognition accuracy is higher for high dimensionality of the proposed method and outperforms the DDL methods at dimensionality 4000 and 12000. In Figure 5 and Figure 6 are shown the recognition accuracy and the expected loss measured as $E[\|x_c^k - y_c^k\|_2^2] = E[1/4x^*_c^k - y_c^k, k, t, M]\$ as a linear function of the discrimination power evaluated at transform dimension $M = \{100, 500, 1500, 4000\}$ and $M = \{1000, 4000, 6000, 12000\}$. It is interesting to highlight that as the discrimination power at different transform dimension increases it also increases the accuracy of recognition. Moreover, the results on these two data sets show that this increase is approximately linear. On the other hand the expected loss decreases as the discrimination power at different transform dimensions increases.

4 CONCLUSION

This paper presented a novel approach for learning discriminative and sparse representations. A novel discriminative prior was proposed and the properties of the models with the prior were investigated. A low complexity learning algorithm was presented. The preliminary results w.r.t. the introduced measures and the recognition accuracy on the used databases showed promising performance. We showed that it is possible to increase the discrimination power with information loss. Moreover, we highlight that when expanding to high dimensional space with non-linear transforms the fact that during learning we used a small number of data samples from the same database and that in the data there is high variability. The conditioning number and the expected coherence for the learned transforms for all the databases at different transform dimensions $M \in Q = \{100, 1150, 2200, 3250, 4300, 5350, 6400, 7450, 8500, 9500, 10600, 11650, 12700, 13750, 14800, 15850, 16900, 17950, 19000 \}$ are shown in Figure 3. We see that the value of both the conditioning number and the coherence is reducing and it converges to common values. This confirms the effectiveness of the conditioning and the coherence constraints. The similarity concentrations $C_1 = D^{B_M}_t(Y)$ and $C^2 = D^{B_M}_t(Y)$, their ratio $C_1/C^2$ and the discrimination power log($C_1/C^2$) = $T^2$ for a subsets of the used databases after applying a non-linear transform having transform dimensions $M \in Q$ is shown in Figure 4. We can see similar behavior as previous, that is, $C_1$ and $C^2$ are decreasing, but, the slope of decrease for $C^2$ is stronger. Therefore, the discrimination power increases as the transform dimension increases.

The proposed method is compared with DLSI (Ramirez et al., 2010), FDDL (Yang et al., 2011a), COPAR (Vu et al., 2015) and LRSSDL (Vu & Monga, 2016). Half of the data samples from the data set Extended YALE B, sampled at random are used for learning and the remaining other half are used for evaluation. Considering the MNIST database the training set is used for learning and the test set is used for both evaluating the discrimination power and the recognition accuracy. We compute both the discrimination power and the recognition accuracy on a subsets from the test sets. The dictionary size (transform dimension $M$) is set to be equal to $\{150, 75, 1515, 3825, 570, 150, 300\}$ for the used databases, respectively, in all of the comparing algorithms. The discrimination power of the comparing methods is denoted as $T^{DLSSI}$, $T^{FDDL}$, $T^{COPAR}$ and $T^{LRSSDL}$. The recognition results for the methods DLSI, FDDL, COPAR and LRSSDL on the data sets Extended YALE B and MNIST were not computed here, rather we use the best reported result form the respective papers (Ramirez et al., 2010), (Yang et al., 2011a), (Vu et al., 2015) and (Vu & Monga, 2016). Considering the proposed algorithm the non-linear transform was learned for the transform dimensions $M = \{100, 500, 1500, 4000\}$ and $M = \{1000, 4000, 6000, 12000\}$, respectively, for the used data sets. After the transform was learned, the transform data samples were computed for the respective training and test sets. Then, the transform training data samples were used as features to learn a linear SVM classifier in one-against-all regime. The results are shown in Table 2(a), b) and c). The discrimination power of the proposed non-linear transform is higher that the discrimination power of the comparing methods. The recognition accuracy is higher for high dimensionality of the proposed method and outperforms the DDL methods at dimensionality 4000 and 12000. In Figure 5 and Figure 6 are shown the recognition accuracy and the expected loss measured as $E[\|x_c^k - y_c^k\|_2^2] = E[1/4x^*_c^k - y_c^k, k, t, M]\$ as a linear function of the discrimination power evaluated at transform dimension $M = \{100, 500, 1500, 4000\}$ and $M = \{1000, 4000, 6000, 12000\}$. It is interesting to highlight that as the discrimination power at different transform dimension increases it also increases the accuracy of recognition. Moreover, the results on these two data sets show that this increase is approximately linear. On the other hand the expected loss decreases as the discrimination power at different transform dimensions increases.
APPENDIX A.

A.1 THE GLOBAL OPTIMAL SOLUTION

Given $X$ and the current estimate of $Y$, the estimate of the transform is a solution to the following problem

$$
\min_A \|AX - Y\|_2^2 + \frac{\lambda_2}{2} \|A\|_F^2 + \frac{\lambda_3}{2} \|AA^T - I\|_F^2 - \lambda_4 \log |\det A^T A|.
$$

(18)

**Theorem 2** (global optimal solution) Given $X \in \mathbb{R}^{N \times CK}$ and $Y \in \mathbb{R}^{M \times CK}$, if and only if the joint decomposition

$$
XX^T = U_X \Sigma_X^2 U_X^T
$$

$$
XY^T = U_X \Sigma_{XY} V_{XY}^T,
$$

(19)

exists, where $U_X \in \mathbb{R}^{N \times N}$ is orthonormal, $V_{XY} \in \mathbb{R}^{M \times N}$ is per columns orthonormal and $\Sigma_X, \Sigma_{XY} \in \mathbb{R}^{N \times N}$ are diagonal matrices with positive diagonal elements, then (18) has a global minimum as

$$
A = V_{XY} \Sigma_A \Sigma_A^{-1} U_X^T,
$$

(20)

$\Sigma_A(n, n) = \sigma_A(n), \forall n, \sigma_A(n) \geq 0$ and $\sigma_A(n)$ are positive solutions to

$$
\frac{\lambda_3}{\sigma_X^2(n)} \sigma_A^3(n) + \frac{\lambda_2}{\sigma_X^2(n)} \sigma_A^2(n) - \frac{\sigma_{XY}(n)}{\sigma_X(n)} \sigma_A(n) - 2\lambda_4 \log \frac{\sigma_X(n)}{\sigma_A(n)} = 0.
$$

(21)

**Proof of Theorem 2** Consider the equivalent trace form of (18)

$$
\min_{A} \text{Tr}\{(AX - Y)^TAX - Y\} + \lambda_2 \text{Tr}\{A^TA\} + \lambda_3 \text{Tr}\{(AA^T - I)(AA^T - I)\} - \lambda_4 \log |A^TA|.
$$

(22)

Note that since $\lambda_2 \geq 0$, $XX^T + \lambda_2 I$ is a symmetric positive definite matrix with all eigenvalues non-negative, therefore it decomposes as

$$
U_X \Sigma_X^2 U_X^T = U_X \Sigma_X U_X^T \Sigma_X U_X^T = XX^T + \lambda_2 I.
$$

(23)

Let

$$
A = BD, D = U_X \Sigma_A^{-1} U_X^T,
$$

(24)

Define

$$
g_1 = BDXY^T, g_2 = BB^T, \quad g_3 = (BDD^T B^T)(BDD^T B^T)^T
$$

$$
g_4 = (BDD^T B^T), \quad g_5 = \log |\det BDD^T B^T|.
$$

(25)

Then (18) equivalently is

$$
\min_B -\text{Tr}\{g_1\} + \text{Tr}\{g_2\} + \lambda_3 \text{Tr}\{g_3 - g_4\} - \lambda_4 g_5.
$$

(26)

Assume that $B$ decomposes as

$$
U_B \Sigma_B V_B^T
$$

(27)

where $\Sigma_B$ is a diagonal matrix with positive diagonal elements, $U_B$ is column orthogonal and $V_B$ is orthogonal square matrix. Moreover, let the following decomposition on $XY^T$ exists

$$
XY^T = U_X \Sigma_X V_{XY}^T, \text{ and substitute as } U_B = V_{XY}, V_B = U_X,
$$

(28)

then

$$
\text{Tr}\{g_1\} = \text{Tr}\{U_B \Sigma_B V_B^T U_X \Sigma_X^{-1} U_X^T \Sigma_X \Sigma_X^{-1} U_X^T \Sigma_X \Sigma_X^{-1} U_X^T\} = \text{Tr}\{\Sigma_B \Sigma_B^{-1} \Sigma_X \Sigma_X^{-1} \Sigma_X \Sigma_X^{-1} \Sigma_X \Sigma_X^{-1} \Sigma_X\}.
$$

(29)

The term

$$
\text{Tr}\{g_2\} = \{BB^T\} = \text{Tr}\{(U_B \Sigma_B V_B^T)(U_B \Sigma_B V_B^T)^T\} = \text{Tr}\{\Sigma_B^2\},
$$

(30)

and

$$
\text{Tr}\{g_4\} = \{BDD^T B^T\} = \text{Tr}\{\Sigma_B \Sigma_X^{-1} \Sigma_X \Sigma_X^{-1} \Sigma_B\} = \text{Tr}\{\Sigma_X^{-1} \Sigma_B^2\}.
$$

(31)

$$
\text{Tr}\{g_5\} = (BDD^T B^T)(BDD^T B^T)^T = \text{Tr}\{\Sigma_X^{-1} \Sigma_B^2\},
$$

(32)

Under review as a conference paper at ICLR 2018
\[ g_s = \log |\det A^T| = \log |\det D^T B^T B| = \log |\det U_X \Sigma_X^{-1} \Sigma_B^2 \Sigma_X^{-1} U_X^T| = \log |\det \Sigma_X^{-2} \Sigma_B^2| \]

Finally, (18) is reduced to

\[
\min_{\sigma_B(n)} \sum_{n=1}^{N} \frac{\lambda_3}{\sigma_X(n)} \sigma_B^4(n) + \frac{2 \lambda_3}{\sigma_X(n)} \sigma_B^2(n) - \frac{\sigma_X(n)}{\sigma_B(n)} \sigma_B(n) - 2 \lambda_4 \log \frac{\sigma_B(n)}{\sigma_A(n)},
\]

equalling to zero the first order derivative of the objective (33) w.r.t. \( \sigma_B(n) \) and multipling by \( \sigma_B(n) \) gives

\[
4 \frac{\lambda_3}{\sigma_X(n)} \sigma_B^4(n) + 2 \frac{\sigma_X(n)}{\sigma_B(n)} \frac{\sigma_X(n)}{\sigma_B(n)} \sigma_B(n) - 2 \lambda_4 = 0
\]

A closed form solution to (34) exists and depends on the discriminant of the quartic polynomial. Moreover, since \( 4 \frac{\lambda_3}{\sigma_X(n)} \) is positive a global minimum to (18) exists if and only if the decomposition (19) exists. \( \square \)

A.2 THE \( \epsilon \)-CLOSE CLOSED FORM APPROXIMATION

Consider the equivalent trace form (26) of (18) and note that since \( \lambda_2 \geq 0 \), \( XX^T + \lambda_2 I \) is a symmetric positive definite matrix with all eigenvalues non-negative, therefore it decomposes as

\[
U_X \Sigma_X^2 U_X^T = U_X \Sigma_X U_X^T U_X \Sigma_X U_X^T = XX^T + \lambda_2 I.
\]

Let the following decomposition exists

\[
U_{U_X XY} \Sigma_{U_X XY} V_{U_X XY}^T = U_X XY^T.
\]

Define

\[
A = BD, \quad D = U_X \Sigma_X^{-1} U_X^T
\]

Assume that \( B \) decomposes as \( U_B \Sigma_B V_B^T = B \), where \( \Sigma_B \) is a diagonal matrix with positive diagonal elements, \( U_B \) is column orthogonal and \( V_B \) is orthogonal square matrix and let

\[
U_B = (U_{U_X XY} V_{U_X XY}^T)^T, \quad V_B = U_X,
\]

then

\[
\text{Tr}\{AXY^T\} = \text{Tr}\{BU_X \Sigma_X^{-1} U_X^T Y^T\} = \text{Tr}\{V_{U_X XY} U_{U_X XY}^T \Sigma_B \Sigma_X^{-1} U_{U_X XY} U_{U_X XY} \Sigma_{U_X XY}^T\}.
\]

Consider the decomposition \( U_B \Sigma_B V_B^T \) of \( B \), use [Mirskey 1959] and [Neumann 1937] and note that

\[
\min_{\Sigma_B} \max_{U_B, V_B} \text{Tr}\{U_B \Sigma_B V_B^T U_X \Sigma_X^{-1} U_X^T X Y^T\} \leq \min_{\Sigma_B} \text{Tr}\{\Sigma_B \Sigma_X^{-1} \Sigma_B\},
\]

where \( \Sigma_B \) is a diagonal matrix, having diagonal elements \( \Sigma_B(n, n) = \sigma_B(n) = T(n, n), \forall n \in N \) and \( T = U_{U_X XY} \Sigma_{U_X XY} V_{U_X XY}^T \).

Note that the term \( \text{Tr}\{(A)(A)^T\} = \text{Tr}\{(BB)^T\} = \text{Tr}\{B^T B\} \) and as in the Appendix subsection A.1 \( \text{Tr}\{(A)(A)^T\} = \text{Tr}\{BDD^T B^T\} = \text{Tr}\{\Sigma_B^2 \Sigma_X^{-2}\} \) and \( \text{log} |\det A| = \text{log} |\det D| = \text{log} |\det D^T B^T B| = \text{log} |\det U_X \Sigma_X^{-1} \Sigma_B^2 \Sigma_X^{-1} U_X^T| = \text{log} |\det \Sigma_X^{-2} \Sigma_B^2| \).

Finally, the approximation of (18) using the bound (41) is reduced to

\[
\min_{\sigma_B(n)} \sum_{n=1}^{N} \frac{\lambda_3}{\sigma_X(n)} \sigma_B^4(n) + \frac{2 \lambda_3}{\sigma_X(n)} \sigma_B^2(n) - \frac{\sigma_X(n)}{\sigma_B(n)} \sigma_B(n) - 2 \lambda_4 \log \frac{\sigma_B(n)}{\sigma_A(n)},
\]

equalling to zero the first order derivative of the objective (42) w.r.t. \( \sigma_B(n) \) and multipling by \( \sigma_B(n) \) gives

\[
4 \frac{\lambda_3}{\sigma_X(n)} \sigma_B^4(n) + 2 \frac{\sigma_X(n)}{\sigma_B(n)} \frac{\sigma_X(n)}{\sigma_B(n)} \sigma_B(n) - 2 \lambda_4 = 0.
\]
A closed form solution to (43) exists and depends on the discriminant of the quartic polynomial. Moreover, since \( 4 - \frac{\lambda_0}{\sigma_X^2} \) is positive a global minimum to (45) exists. Therefore, having the decomposition \( U_B \Sigma_B V_B^T = B \), the substitutions \( U_B = (U_X Y V_{U_X Y}^T)^T \) and \( V_B = U_X \) with the solution of (45) gives the \( \epsilon \)-close closed form approximative solution to problem (18) as

\[
A = V_{U_X Y} U_{U_X Y}^T \Sigma_B \Sigma_X^{-1} U_X^T,
\]

where the bound (41) implies that the \( \epsilon \)-close closed form approximative solution is a lower bound to the solution of (18) \( \square \).

**APPENDIX B.**

Let \( y_{c,k} = y_{c,k}^+ + y_{c,k}^- \), with \( y_{c,k}^+ \in \mathbb{R}_M^+ \) and \( y_{c,k}^- \in \mathbb{R}_M^- \). Consider the similarity concentration measure

\[
D_{c,k}^T(X) = \sum_{c_1 \neq c_2} \sum_{k_1, k_2} \| y_{c_1,k_1}^+ \odot y_{c_2,k_2}^+ \|_1 + \sum_{c_1 \neq c_2} \sum_{k_1, k_2} \| y_{c_1,k_1}^- \odot y_{c_2,k_2}^- \|_1 = \sum_{c_1 \neq c_2} \sum_{k_1, k_2} \| y_{c_1,k_1}^+ \odot y_{c_2,k_2}^+ \|_1 + \sum_{c_1 \neq c_2} \sum_{k_1, k_2} \| y_{c_1,k_1}^- \odot y_{c_2,k_2}^- \|_1 = \sum_{c_1 \neq c_2} \sum_{k_1, k_2} \| y_{c_1,k_1}^+ \|_1 \| y_{c_2,k_2}^+ \|_1 + \sum_{c_1 \neq c_2} \sum_{k_1, k_2} \| y_{c_1,k_1}^- \|_1 \| y_{c_2,k_2}^- \|_1
\]

\[
\| y_{c_1,k_1}^+ \|_1 \| y_{c_2,k_2}^+ \|_1 + \sum_{c_1 \neq c_2} \sum_{k_1, k_2} \| y_{c_1,k_1}^- \|_1 \| y_{c_2,k_2}^- \|_1 = \sum_{c_1 \neq c_2} \sum_{k_1, k_2} \| y_{c_1,k_1}^+ \|_1 \| y_{c_2,k_2}^+ \|_1 + \sum_{c_1 \neq c_2} \sum_{k_1, k_2} \| y_{c_1,k_1}^- \|_1 \| y_{c_2,k_2}^- \|_1
\]

where \( g^+ = \sum_{c_1 \neq c_2} \sum_{k_1, k_2} y_{c_1,k_1}^+ \odot y_{c_2,k_2}^+ \) and \( g^- = \sum_{c_1 \neq c_2} \sum_{k_1, k_2} y_{c_1,k_1}^- \odot y_{c_2,k_2}^- \) and we abuse notation by denoting \( |y_{c_1,k_1}| \) as the vector whose elements are the absolute values of the elements in \( y_{c_1,k_1} \).

Note that

\[
S_{c,k}^T(X) = \sum_{c_1 \neq c_2} \sum_{k_1, k_2} \| y_{c_1,k_1} \odot y_{c_2,k_2} \|_2^2 = \sum_{c_1 \neq c_2} \sum_{k_1, k_2} (y_{c_1,k_1} \odot y_{c_2,k_2})^T (y_{c_2,k_2} \odot y_{c_2,k_2}) = (y_{c_1,k_1} \odot y_{c_1,k_1})^T (\sum_{c_1 \neq c_2} \sum_{k_1, k_2} y_{c_2,k_2}^+ \odot y_{c_2,k_2}^+ )
\]

where \( s_c = \sum_{c_1 \neq c_2} \sum_{k_1, k_2} y_{c_1,k_1}^+ \odot y_{c_2,k_2}^+ \). Denote \( q_{c_1,k_1} = Ax_{c_1,k_1} \) and consider the problem

\[
\min_{y_{c_1,k_1}} \| q_{c_1,k_1} - y_{c_1,k_1} \|_2^2 + \lambda_0 (y_{c_1,k_1}^T g^+ + (y_{c_1,k_1}^T g^-) + (y_{c_1,k_1} \odot y_{c_1,k_1})^T s_c) + \lambda_1 \| y_{c_1,k_1} \|_1,
\]

by taking the first order derivative w.r.t. \( y_{c_1,k_1} \) we have that

\[
(y_{c_1,k_1} - q_{c_1,k_1}) + \lambda_0 (\operatorname{sign}(y_{c_1,k_1}) \odot g^+ + \operatorname{sign}(y_{c_1,k_1}) \odot g^- + y_{c_1,k_1} \odot s_c) + \lambda_1 \operatorname{sign}(y_{c_1,k_1}) = 0,
\]

take sign magnitude decomposition of \( y_{c_1,k_1} = \operatorname{sign}(y_{c_1,k_1}) \odot |y_{c_1,k_1}| \) then we have

\[
\operatorname{sign}(y_{c_1,k_1}) \odot |y_{c_1,k_1}| \odot (1 + 2 \lambda_0 s_c) - \operatorname{sign}(q_{c_1,k_1}) \odot |q_{c_1,k_1}| + \lambda_0 (\operatorname{sign}(y_{c_1,k_1}) \odot g^+ + \operatorname{sign}(y_{c_1,k_1}) \odot g^-) + \lambda_1 \operatorname{sign}(y_{c_1,k_1}) = 0.
\]

Let the sign of \( y_{c_1,k_1} \), i.e. \( \operatorname{sign}(y_{c_1,k_1}) \) be equal to the sign of \( \operatorname{sign}(q_{c_1,k_1}) \), and Hadamard multiply from the left side by \( \operatorname{sign}(q_{c_1,k_1}) \) then we have

\[
|y_{c_1,k_1}| \odot (1 + 2 \lambda_0 s_c) - |q_{c_1,k_1}| + \lambda_0 (\operatorname{sign}(q_{c_1,k_1}) \odot \operatorname{sign}(y_{c_1,k_1}) \odot g^+ + \operatorname{sign}(q_{c_1,k_1}) \odot \operatorname{sign}(y_{c_1,k_1}) \odot g^-) + \lambda_1 1 = 0.
\]
note that \( \text{sign}(q^{+}_{c1,k1}) \circ \text{sign}(q^{+}_{c1,k1}) = \text{sign}(q^{+}_{c1,k1}) \) and that \( \text{sign}(q_{c1,k1}) \circ \text{sign}(q^{+}_{c1,k1}) = \text{sign}(-q^{-}_{c1,k1}) \), therefore we have

\[
|y_{c1,k1}| \circ (1 + 2\lambda_0 s_c) = |q_{c1,k1}| - \lambda_0(\text{sign}(q^{+}_{c1,k1}) \circ g^{+} + \text{sign}(-q^{-}_{c1,k1}) \circ g^{-}) - \lambda_1 1. \tag{51}
\]

since the magnitude might be only positive we have that \( |y_{c1,k1}| \circ (1 + 2\lambda_0 s_c) = \max(|q_{c1,k1}| - \lambda_0(\text{sign}(q^{+}_{c1,k1}) \circ g^{+} + \text{sign}(-q^{-}_{c1,k1}) \circ g^{-}) - \lambda_1 1, 0) \). Denote \( g = (\text{sign}(q^{+}_{c1,k1}) \circ g^{+} + \text{sign}(-q^{-}_{c1,k1}) \circ g^{-}) \) then the closed form solution to (47) is:

\[
y_{c1,k1} = \text{sign}(y_{c1,k1}) \circ \max(|q_{c1,k1}| - \lambda_0 g - \lambda_1 1, 0) \circ (1 + \lambda_0 s_c). \tag{52}
\]
which completes the proof \( \square \)

### Appendix C. Sensitivity Analysis and Interpretations

The similarity concentration measure provides possibility to measure the discriminative properties, their deviation, increase (or decrease) and the corresponding relations between different non-linear transform models across one domain or different domains, thereby quantifying their quality w.r.t. the discriminative properties.

#### C.1 Sensitivity Analysis w.r.t. the Similarity Concentration Measures

An illustration about the definition of discriminative power given by a diagram is shown in Figure 7.

To measure the ability for an increase in discriminative properties by a non-linear transform we first have to define a notion for the discriminative properties on a data set under different non-linear transform models. Therefore, first we introduce the "special" base models and then analyze the properties of the similarity concentration measures under the change in model parameter and the relation between the base models and the proposed non-linear transform model defined by a parameter set \( \mathcal{P} = \{A \in \mathbb{R}^{M \times N}, \tau \in \mathbb{R}^{M}\} \).

Any data set \( X \) in the original domain might have a transform model with parameters \( B^N = \{A_o \in \mathbb{R}^{N \times N}, \tau = 0 \in \mathbb{R}^{N}\} \), if \( A_o = 1 \in D^N \) we refer to it as a base original model. Similarly as in the original domain, any data set \( Y \) in the transform domain might have a transform model with parameters \( B^M = \{A_t \in \mathbb{R}^{M \times M}, \tau = 0 \in \mathbb{R}^{M}\} \), if \( A_t = 1 \in D^M \) we refer to it as a base transform model. Any base model, defined either in the original domain \( B^N \) or in the transform domain \( B^M \), has domain equal to the co-domain, since \( x_{c,k} = T^N (x_{c,k}) \) and \( y_{c,k} = T^M (y_{c,k}) \) holds trivially, for the respective sets of parameters \( B^N = \{A_o = 1 \in D^N, \tau = 0 \in \mathbb{R}^{N}\} \) and \( B^M = \{A_t = 1 \in D^M, \tau = 0 \in \mathbb{R}^{M}\} \). This is illustrated with a diagram shown in Figure 8.

A base original model provides a possibility to compare it with any other non-linear transform model with parameters \( \mathcal{P} = \{A \in \mathbb{R}^{M \times N}, \tau \in \mathbb{R}^{M}\} \). Additionally, note that for \( B^M = \{A_t = 1 \in D^M, \tau = 0 \in \mathbb{R}^{M}\} \) we have that \( D^M_{1,c}(Y) = D^P_{1,c}(X) \) and that \( D^M_{1}(Y) = D^P_{1}(X) \). It implies

\[
\begin{align*}
X & \xrightarrow{T^P} Y \\
\log \frac{D^N_{1,c}(X)}{D^N_{1}(X) + \epsilon} & \xrightarrow{\mathcal{I}^o} \log \frac{D^P_{1,c}(X)}{D^P_{1}(X) + \epsilon} \\
\end{align*}
\]

Figure 7: The relation for the definition of the discrimination power in the original and the transform domain under the base models \( B^N \) and \( B^M \).
The proof is given in Appendix C.2.

Figure 8: The original and the transform domains under a non-linear transforms with a set of parameters $B^N$, $P$ and $B^M$. note that for $T^{B^N}$ and $T^{B^M}$ the original and the transform domains are the same.

that the similarity concentrations can be analyzed as a function in the original domain under model $P$ or in the transform domain under model $B^M$. The main relations considering the preservation of change in the similarity concentration between two models, defined not necessary in the same domain are stated by  Lemma 1.

**Lemma 1:** The non-linear transform model (2) totally preserves the information in the change for the similarity concentration for a data set $X$ w.r.t. a small change in the parameters of the models $B^N$, $B^M$ and $P$ if $\|\delta_o\|_*$ = 0 and $\|\delta_t\|_*$ = 0 as

\[
(R1) : A \left( \frac{\partial D_{t,c}^{B^N}(X)}{\partial A_o} |_{A_o=1} + \frac{\partial D_{t,c}^{B^N}(X)}{\partial A_o} |_{A_o=1} \right) = \frac{\partial D_{t,c}^{P}(X)}{\partial A} + \frac{\partial D_{t,c}^{P}(X)}{\partial A} + \delta_o
\]

\[
(R2) : A \left( \frac{\partial D_{t,c}^{P}(X)}{\partial A} + \frac{\partial D_{t,c}^{P}(X)}{\partial A} \right)^T = \frac{\partial D_{t,c}^{B^M}(Y)}{\partial A_t} |_{A_t=1} + \frac{\partial D_{t,c}^{B^M}(Y)}{\partial A_t} |_{A_t=1} + \delta_t
\]

\[
(R3) : A \delta_o^T = \frac{\partial D_{t,c}^{B^M}(Z)}{\partial A_t} |_{A_t=1} + \frac{\partial D_{t,c}^{B^M}(Z)}{\partial A_t} |_{A_t=1} + \delta_t^T
\]

where

\[
\frac{\partial D_{t,c}^{B^N}(X)}{\partial A} + \frac{\partial D_{t,c}^{B^N}(X)}{\partial A} = \sum_{c,e_l} \frac{\sum_{k,k_1} Y_{c,k} X_{c,k_1}^T + Y_{c,k} X_{c,k_1}^T}{k,k_1}
\]

\[
\frac{\partial D_{t,c}^{B^M}(Y)}{\partial A_t} |_{A_t=1} + \frac{\partial D_{t,c}^{B^M}(Y)}{\partial A_t} |_{A_t=1} = \sum_{c,e_l} \frac{\sum_{k,k_1} Y_{c,k} Y_{c,k_1}^T + Y_{c,k} Y_{c,k_1}^T}{k,k_1}
\]

\[
\frac{\partial D_{t,c}^{B^M}(Z)}{\partial A_t} |_{A_t=1} + \frac{\partial D_{t,c}^{B^M}(Z)}{\partial A_t} |_{A_t=1} = \sum_{c,e_l} \frac{\sum_{k,k_1} Z_{c,k} Z_{c,k_1}^T + Z_{c,k} Z_{c,k_1}^T}{k,k_1}
\]

\[
\frac{\partial D_{t,c}^{B^N}(X)}{\partial A_o} |_{A_o=1} = \sum_{c,k,k_1} X_{c,k} X_{c,k_1}^T + X_{c,k} X_{c,k_1}^T
\]

\[
\delta_o = \sum_{c,e_l} \frac{\sum_{k,k_1} Z_{c,k} X_{c,k_1}^T + Z_{c,k} X_{c,k_1}^T}{k,k_1}
\]

\[
\delta_t = \sum_{c,e_l} \frac{\sum_{k,k_1} Z_{c,k} Y_{c,k_1}^T + Z_{c,k} Y_{c,k_1}^T}{k,k_1}
\]

The proof is given in Appendix C.2.
The terms $z_{c,k}$ represent the non-linear transform error vectors that appear in the model $A x_{c,k} = y_{c,k} + z_{c,k}$ as a result of applying an element-wise non-linear operation $H_{\tau}$ to $A x_{c,k}$, i.e., $y_{c,k} = H_{\tau}(A x_{c,k})$. As an example in the sparsifying transform model $z_{c,k}$ is the "loss of information", that is the information about the values of the elements in $A x_{c,k}$ that are discarded. The terms $\delta_o$ and $\delta_i$ correlate the errors $z_{c,k}$ with the original data $x_{c,k}$ and transform data $y_{c,k}$, respectively. Note that if there is no loss of information (in the earlier example it means that there is no thresholding and just a simple linear transform model is used) then $\delta_o = 0$ and $\delta_i = 0$. Moreover, $\delta_o$ and $\delta_i$ bear important information about the discriminative properties in the transform domain.

The terms $\frac{\partial D_{\ell_1}(X)}{\partial A}$ and $\frac{\partial D_{\ell_1}(X)}{\partial A}$ represent the change of the similarity concentrations under infinitesimally small change of the parameter $A$ from the model $P$. The terms $\frac{\partial D_{\ell_1}(X)}{\partial A}|_{A_o=1}$ and $\frac{\partial D_{\ell_1}(X)}{\partial A}|_{A_o=1}$ have dual interpretation. Assuming metric $A_o = I$, then the first one is considered as a change of the similarity concentrations under infinitesimally small change of the space metric, or equivalently under small metric perturbation. Conversely, assuming the data samples are distributed under a Gaussian distribution with parameters identity covariance matrix and zero mean, i.e., $x_{c,k} \sim \mathcal{N}(\mu = 0, \Sigma = I)$, then $\frac{\partial D_{\ell_1}(X)}{\partial A}|_{A_o=1}$ and $\frac{\partial D_{\ell_1}(X)}{\partial A}|_{A_o=1}$ represent the change of the similarity concentrations under small change in the assumption away from a Gaussian distribution.

Equation (R1) relates the base transform for the original domain $B^N$ with any arbitrary transform defined in the original domain $P$. The relation (R2) is a result about the preservation of change in the similarity concentration between two models $B^M$ and $P$ defined on two different domains. Whereas (R3) gives the preservation of change in the similarity concentration between the error in the transform domain.

The next result highlights the relationship between: the linear projection (by the linear map $A$ that appears in the model $P$) of the change in the similarity concentration under the model $B^N$ in the original domain and the change of the similarity concentration under the model $B^M$ in the transform domain.

This relation exists independently for $\frac{\partial D_{\ell_1}(X)}{\partial A}|_{A_o=1}$ and $\frac{\partial D_{\ell_1}(X)}{\partial A}|_{A_o=1}$, nevertheless, we will define the summarized and the independent versions. Therefore, first we define $\frac{\partial J_{\ell_1}(X)}{\partial A}|_{A_o=1} = \frac{\partial D_{\ell_1}(X)}{\partial A}|_{A_o=1} + \frac{\partial D_{\ell_1}(X)}{\partial A}|_{A_o=1}$ and $\frac{\partial J_{\ell_1}(Y)}{\partial A}|_{A_o=1} = \frac{\partial D_{\ell_1}(Y)}{\partial A}|_{A_o=1} + \frac{\partial D_{\ell_1}(Y)}{\partial A}|_{A_o=1}$ and rewrite (R1) as $A \frac{\partial J_{\ell_1}(X)}{\partial A}|_{A_o=1} - \delta_o = \frac{\partial J_{\ell_1}(Y)}{\partial A}|_{A_o=1} + \frac{\partial D_{\ell_1}(X)}{\partial A} + \frac{\partial D_{\ell_1}(X)}{\partial A}$, then replace $\frac{\partial D_{\ell_1}(X)}{\partial A}$ + $\frac{\partial D_{\ell_1}(X)}{\partial A}$ in (R2) by the same term in (R1), use (R3), reorder and we have the following result.

**Lemma 2:** For fixed $\tau$ any non-linear transform model (2) preserves the information in the change of similarity concentrations w.r.t. a small change in $A$ by

\begin{align}
\text{(R4): } A \frac{\partial J_{\ell_1}(X)}{\partial A}|_{A_o=1} A^T & = \frac{\partial J_{\ell_1}(V)}{\partial A}|_{A_o=1} = \frac{\partial J_{\ell_1}(Y)}{\partial A}|_{A_o=1} + \xi_e + \xi, \\
\text{(R5): } A \frac{\partial D_{\ell_1,c}(X)}{\partial A}|_{A_o=1} A^T & = \frac{\partial D_{\ell_1,c}(V)}{\partial A}|_{A_o=1} = \frac{\partial D_{\ell_1,c}(Y)}{\partial A}|_{A_o=1} + \xi_e + \xi, \\
\text{(R6): } A \frac{\partial D_{\ell_1}(X)}{\partial A}|_{A_o=1} A^T & = \frac{\partial D_{\ell_1}(V)}{\partial A}|_{A_o=1} = \frac{\partial D_{\ell_1}(Y)}{\partial A}|_{A_o=1} + \xi, 
\end{align}

where

$$V = AX$$

\begin{align}
\xi_e + \xi & = \frac{\partial J_{\ell_1}(Z)}{\partial A}|_{A_o=1} + \frac{\partial J_{\ell_1}(Y; Z)}{\partial A}|_{A_o=1} \\
\frac{\partial J_{\ell_1}(Z)}{\partial A}|_{A_o=1} & = \frac{\partial D_{\ell_1,c}(Z)}{\partial A}|_{A_o=1} + \frac{\partial D_{\ell_1}(Z)}{\partial A}|_{A_o=1} \\
\frac{\partial J_{\ell_1}(Y; Z)}{\partial A}|_{A_o=1} & = \delta_i + \delta_i^T 
\end{align}
The expressions \((R4)\) actually relate the metric in the original domain under the model \(B^N\) to the induced metric in the transform domain for the model \(B^M\), with induction done by the model \(\mathcal{P}\) with parameter set \(\{A \in \mathbb{R}^{M \times N}, \tau \in \mathbb{R}^M\}\). Moreover, the model \(\mathcal{P}\) might describe a transform domain with a non-smooth manifold. Since the manifolds of the original and the transform domain under the models \(B^N\) and \(B^M\) are smooth the analysis of their relations reveals insights about the relation between the manifolds under the models \(B^N\) and \(\mathcal{P}\). The terms \(\frac{\partial J_{I_1}(Z)}{\partial A_o}|_{A_o=1}\) and \(\frac{\partial J_{I_1}(Y,Z)}{\partial A_o}|_{A_o=1}\) carry out the information about the breaks and the discontinuities of the regularity and smoothness in the manifold induced by the model \(\mathcal{P}\). Also we note that \(\frac{\partial J_{I_1}(X)}{\partial A_o}|_{A_o=1} = 4\frac{\partial J_{I_1}(X)}{\partial A_o}|_{A_o=1}\) and \(\frac{\partial J_{I_1}(Y)}{\partial A_o}|_{A_o=1} = 4\frac{\partial J_{I_1}(Y)}{\partial A_o}|_{A_o=1}\) therefore, \(\frac{\partial J_{I_1}(X)}{\partial A_o}|_{A_o=1} = 4\frac{\partial J_{I_1}(X)}{\partial A_o}|_{A_o=1}\). Since the manifolds of the original and the transform domain under the models \(B^N\) and \(\mathcal{P}\) change of coordinates on a manifold, where the intrinsic properties of curvature remain unchanged.

**APPENDIX C.2**

Note that for the model \(\mathcal{P}^t\) we have that

\[
y = \mathcal{T}(Ax) = \max(Ax - \tau, 0) - \max(-Ax - \tau, 0),
\]

\[
q = \mathcal{T}(Ax) = \max(Ag - \tau, 0) - \max(-Ag - \tau, 0),
\]

(57)

since

\[
\text{sign}(a) \max(|a| - b, 0) = \max(a - b, 0) - \max(-a - b, 0),
\]

(58)

The first order derivative of the divergence \(D_{\ell_1}^{\mathcal{P}^t}(x; g)\) w.r.t. the parameter \(A\) is:

\[
\frac{\partial D_{\ell_1}^{\mathcal{P}^t}(x; g)}{\partial A} = \frac{\partial (\max(Ax - \tau, 0)^T \max(Ag - \tau, 0))}{\partial A} + \frac{\partial (\max(-Ax - \tau, 0)^T \max(-Ag - \tau, 0))}{\partial A}
\]

(59)

we assume that the threshold parameter \(\tau\) is chosen such that the vector \(|Ax| - \tau\) (or for any other \(q\), the vector \(|Ag| - \tau\) ) has least one non-zero element, then

\[
\frac{\partial (\max(Ax - \tau, 0)^T \max(Ag - \tau, 0))}{\partial A} = \max(Ag - \tau, 0)x^T + \max(Ax - \tau, 0)g^T,
\]

(60)

and

\[
\frac{\partial (\max(-Ax - \tau, 0)^T \max(-Ag - \tau, 0))}{\partial A} = - \max(-Ag - \tau, 0)x^T - \max(-Ax - \tau, 0)g^T,
\]

(61)

combining (60) and (61) we have that

\[
\frac{\partial D_{\ell_1}^{\mathcal{P}^t}(x; g)}{\partial A} = \max(Ag - \tau, 0)x^T + \max(Ax - \tau, 0)g^T - \max(-Ag + \tau, 0)x^T + \max(-Ax + \tau, 0)g^T = qx^T + yg^T
\]

(62)

where

\[
y(m) = \text{sign}(a^m_Tx) \max(|a^m_Tx| - \tau(m), 0)
\]

\[
q(m) = \text{sign}(a^m_Tg) \max(|a^m_Tg| - \tau(m), 0),
\]

(63)

\(\forall m \in \mathcal{M}\).
Similarity, note that for the model $\mathcal{P}_0^\tau = \{A_o, \tau = 0\}$ we have that

$$x_o = T_{\mathcal{P}_0^\tau}^\tau (A_o x) = \begin{cases} \text{sign}(A_o x) \odot \max(|A_o x| - 0, 0) = \\
\max(A_o x - 0, 0) - \max(-A_o x - 0, 0), \end{cases} \text{ (64)}$$

$$g_o = T_{\mathcal{P}_0^\tau}^\tau (A_o g) = \begin{cases} \text{sign}(A_o g) \odot \max(|A_o g| - g, 0) = \\
\max(A_o g - 0, 0) - \max(-A_o g - 0, 0), \end{cases}$$

The first order derivative of the divergence $D_{\ell_1}^{P_0^\tau}(x; g)$ w.r.t $A_o$ is:

$$\frac{\partial D_{\ell_1}^{P_0^\tau}(x; g)}{\partial A_o} = x_o g^T + g_o x^T, \text{ (65)}$$

note that at $A_o = I, \mathcal{P}_0^\tau = B^N$ and we have that

$$\frac{\partial D_{\ell_1}^{P_0^\tau}(x; g)}{\partial A_o} |_{A_o = I} = x g^T + g x^T \text{ (66)}$$

Also for the model $\mathcal{P}_0^\tau = \{A_t, \tau = 0\}$ we have that

$$y_t = T_{\mathcal{P}_0^\tau}^\tau (A_t y) = \begin{cases} \text{sign}(A_t y) \odot \max(|A_t y| - 0, 0) = \\
\max(A_t y - 0, 0) - \max(-A_t y - 0, 0), \end{cases} \text{ (67)}$$

$$q_t = T_{\mathcal{P}_0^\tau}^\tau (A_t q) = \begin{cases} \text{sign}(A_t q) \odot \max(|A_t q| - q, 0) = \\
\max(A_t q - 0, 0) - \max(-A_t q - 0, 0). \end{cases}$$

The first order derivative of the divergence $D_{\ell_1}^{P_0^\tau}(y; q)$ w.r.t $A_t$ is:

$$\frac{\partial D_{\ell_1}^{P_0^\tau}(y; q)}{\partial A_t} = y_t q^T + q_t y^T, \text{ (68)}$$

note that at $A_t = I, \mathcal{P}_0^\tau = B^M$ and we have that

$$\frac{\partial D_{\ell_1}^{P_0^\tau}(y; q)}{\partial A_t} |_{A_t = I} = y q^T + q y^T \text{ (69)}$$

Consider the following

$$Ax = y + z_2/g^T \rightarrow \begin{cases} A x g^T = y g^T + z_1 g^T = \\
A g x^T = q x^T + z_2 x^T, \end{cases} \text{ (70)}$$

$$A(x g^T + g x^T) = y g^T + q x^T + z_1 g^T + z_2 x^T$$

where

$$z_1 = A x - \text{sign}(A x) \max(|A x| - \tau 1, 0)$$
$$z_2 = A g - \text{sign}(A g) \max(|A g| - \tau 1, 0) \text{ (71)}$$

A closer look at (70) reveals us that

$$\frac{\partial D_{\ell_1}^{B^N}(x; g)}{\partial A_o} |_{A_o = I} = \frac{\partial D_{\ell_1}^{P_0^\tau}(x; g)}{\partial A} + \delta_{i=1}^{p_1} z_2 \text{ (72)}$$
where
\[ \delta_o^{z_1,z_2} = z_1g^T + z_2x^T \] (73)

By similar construction applied to the rest of the pairs of data samples we have the result:
\[ A \left( \frac{\partial D_{t,c}^N(X)}{\partial A_o} |_{A_o=1} + \frac{\partial D_{t,c}^N(X)}{\partial A_o} |_{A_o=1} \right) = \frac{\partial D_P^N(X)}{\partial A} + \frac{\partial D_P^N(X)}{\partial A} + \delta_o \] (74)

Note that
\[ \begin{align*}
Ax &= y + z_1/q^T \\
Ag &= q + z_2/y^T \\
A(yq^T + qg^T) &= yq^T + qy^T + z_1q^T + z_2y^T,
\end{align*} \] (75)

where
\[ z_1 = Ax - \text{sign}(Ax) \max(|Ax| - \tau 1, 0) \]
\[ z_2 = Ag - \text{sign}(Ag) \max(|Ag| - \tau 1, 0) \] (76)

A closer look at (75) reveals us
\[ A \left( \frac{\partial D_{t}^N(x; g)}{\partial A} \right)^T = \frac{\partial D_{t,c}^M(Y; q)}{\partial A} |_{A_t=1} + \delta_t \] (77)

where
\[ \delta_t^{z_1,z_2} = z_1q^T + z_2y^T \] (78)

By similar construction applied to the rest of the pairs of data samples we have the result:
\[ A \left( \frac{\partial D_{t,c}^M(X)}{\partial A} + \frac{\partial D_{t,c}^M(X)}{\partial A} \right)^T = \left( \frac{\partial D_{t}^M(Y)}{\partial A_t} |_{A_t=1} + \frac{\partial D_{t}^M(Y)}{\partial A_t} |_{A_t=1} \right) + \delta_t \] (79)

Note that
\[ \begin{align*}
Ax &= y + z_1/z_2^T \\
Ag &= q + z_2/z_1^T \\
A(z_2q^T + qz_1^T) &= yz_2^T + qz_1^T + z_1q^T + z_2y^T,
\end{align*} \] (80)

where
\[ z_1 = Ax - \text{sign}(Ax) \max(|Ax| - \tau 1, 0) \]
\[ z_2 = Ag - \text{sign}(Ag) \max(|Ag| - \tau 1, 0) \] (81)

A closer look at (80) reveals us
\[ A(\delta_t^{z_1,z_2})^T = \frac{\partial D_{t}^M(z_1; z_2)}{\partial A_t} |_{A_t=1} + (\delta_t^{z_1,z_2})^T, \] (82)

where
\[ \delta_t^{z_1,z_2} = z_1q^T + z_2y^T \] (83)

By similar construction applied to the rest of the pairs of data samples we have the result:
\[ A\delta_o^T = \left( \frac{\partial D_{t,c}^M(Z)}{\partial A_t} |_{A_t=1} + \frac{\partial D_{t}^M(Z)}{\partial A_t} |_{A_t=1} \right) + \delta_t \] (84)
Figure 9: The ratio $C1/C2$ of the similarity concentrations $C1 = D_{t_1,c}^B(Y)$ and $C2 = D_{t_1}^B(Y)$ and the discrimination power $\log(C1/C2) = I^t$ for randomly chosen subsets from all of the used databases under a non-linear transform with transform dimension $M = 19000$ and varying thresholding parameter $\tau = \lambda 1$.

APPENDIX D

The result in by Lemma 2 decomposes on the contribution components for similarity and on the contribution components for dissimilarity

$$\frac{\partial J_{t_1}}{\partial A_t} |_{A_t=1} = \frac{\partial J_{t_1}}{\partial A_t} |_{A_t=1}| - \frac{\partial J_{t_1}}{\partial A_t} |_{A_t=1}|_s - \frac{\partial J_{t_1}}{\partial A_t} |_{A_t=1}|_d + \xi c|_s - \xi c|_d + \xi s - \xi d.$$

Moreover, w.r.t. the similarity concentrations we have the following decompositions

$$Tr \left\{ \frac{\partial D_{t_1,c}^B(AX)}{\partial A_t} |_{A_t=1} \right\} = D_{t_1,c}^B(Y) + Tr \{ \xi c|_s \} = D_{t_1,c}^B(X) + Tr \{ \xi c|_s \}$$

$$Tr \left\{ \frac{\partial D_{t_1}^B(AX)}{\partial A_t} |_{A_t=1} \right\} = D_{t_1}^B(Y) + Tr \{ \xi |_s \} = D_{t_1}^B(X) + Tr \{ \xi |_s \},$$

therefore we have the following bounds

$$a: Tr \left\{ A \frac{\partial D_{t_1,c}^N(X)}{\partial A_o} |_{A_o=1} A^T \right\} \leq D_{t_1,c}^P(X) \leq Tr \left\{ A \frac{\partial D_{t_1,c}^B(AX)}{\partial A_t} |_{A_t=1} \right\} = D_{t_1,c}^B(AX)$$

$$b: Tr \left\{ A \frac{\partial D_{t_1}^N(X)}{\partial A_o} |_{A_o=1} A^T \right\} \leq D_{t_1}^P(X) \leq Tr \left\{ A \frac{\partial D_{t_1}^B(AX)}{\partial A_t} |_{A_t=1} \right\} = D_{t_1}^B(AX)$$

(87)

Additionally, we have that $c: \lambda_{\min}(A^T A) Tr \left\{ \frac{\partial D_{t_1}^N(X)}{\partial A_o} |_{A_o=1} \right\} \leq Tr \left\{ A \frac{\partial D_{t_1}^N(X)}{\partial A_o} |_{A_o=1} A^T \right\}$, where $\lambda_{\min}(A^T A)$ is the minimum singular value of the matrix $A^T A$. Taking the logarithm of the ratio $\frac{D_{t_1}^P(X)}{D_{t_1}^B(AX)}$ using the bounds $a$, $b$ and $c$ we arrive at the desired result.

APPENDIX E

The exact steps of the proposed non-linear transform learning are described by Algorithm 1.

APPENDIX F

The ratio $C1/C2$ between the similarity concentrations $C1 = D_{t_1,c}^B(Y)$ and $C2 = D_{t_1}^B(Y)$ and the discrimination power $\log(C1/C2) = I^t$ on subsets of the used databases after applying a non-linear transform with transform dimension $M = 19000$ and varying thresholding parameter $\tau = \lambda 1$ is shown in Figure 9. We used 70 different values for the parameter $\lambda$, sampled uniformly from the interval $\{0, (\max_{1 \leq c \leq C, 1 \leq k \leq K} \max_{1 \leq m \leq M} |a_{m,c,k}|)\}$. The results were obtained using
Algorithm 1: Non-linear transform learning algorithm

Input $X, \lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4$

A $\leftarrow$ initialize

repeat

DISCRIMINATIVE ENCODING closed form solution per data sample

$Y \leftarrow AX$

repeat

for all $c \in C$

$d_c^- \leftarrow \sum_{c \neq c} \sum_k y_{c,k}^-$, $d_c^+ \leftarrow \sum_{c \neq c} \sum_k y_{c,k}^+$ and

$s_c \leftarrow \sum_{c \neq c} \sum_k y_{c,k} \circ y_{c,k}$

end for

for all $c \neq C$ and $\forall k \in K$

$g \leftarrow \text{sign}(\max(Ax_{c,k}, 0)) \odot d_c^+ + \text{sign}(\min(-Ax_{c,k}, 0)) \odot d_c^-$

$y_{c,k} \leftarrow \text{sign}(Ax_{c,k}) \odot \max(Ax_{c,k} - \lambda_0 g + \lambda_1 1, 0) \odot (1 + 2\lambda_0 s_c)$

end for

until convergence

TRANSFORM UPDATE $\varepsilon$-close closed form solution

$U_X \Sigma_X^2 U_X^T \leftarrow XX^T + \lambda_2 I$ and

$U_{XXY} \Sigma_{XXY} U_{XXY}^T \leftarrow U_{XXY}^T X Y^T$

$\min_{\sigma_A(n)} \frac{\lambda}{\sigma_A^2(n)} \sigma_A(n) + \left(\frac{\sigma_X^2(n) - 2\lambda_3}{\sigma_X^2(n)}\right) \sigma_A^2(n) - \frac{\sigma_X^2(n)}{\sigma_X^2(n)} \sigma_A(n) - 2\lambda_4 \log \frac{\sigma_A(n)}{\sigma_X(n)}$

where $\sigma_A(n) \leftarrow T(n, n)$, $T \leftarrow U_{XXY} \Sigma_{XXY} U_{XXY}^T$, $\forall n \in N$

$A \leftarrow V_{XXY} U_{XXY} \Sigma_A \Sigma_X^{-1} U_X^T$

until convergence

Output $A, Y$

Figure 10: The ratio of the similarity concentrations similarity concentrations $C1 = D_{\chi_1}^{BM}(Y)$ and $C2 = D_{\chi_2}^{BM}(Y)$ and the discrimination power $\log(C1/C2) = T^d$ for the Extended Yale B and MNIST databases under non-linear transforms having different transform dimension $M$ and varying thresholding parameter $\tau = \lambda I$.

A non-linear transform learned with one value for parameter $\lambda$ for all the databases. Since all the databases have different variabilities and the amount of available data is different, this result suggest that per different database there should be different optimal values for the parameter $\lambda$. 

23
The results about the ratio \( C_1/C_2 \) between the similarity concentrations \( C_1 = D^M_{\beta_{c1}}(Y) \) and \( C_2 = D^M_{\beta_{c1}}(Y) \) and the discrimination power \( \log(C_1/C_2) = I \) on the Extended-Yale-B and the MNIST data sets after applying a non-linear transform with transform dimensions \( M = \{100, 500, 1500, 4000\} \) and \( M = \{1000, 4000, 6000, 12000\} \), and varying the thresholding parameter \( \tau = \lambda_1 \) are shown in Figure 10. We again used 70 different values for the parameter \( \lambda \), sampled uniformly from the interval \([0, (\max_{1 \leq c \leq C, 1 \leq k \leq K, \max_{1 \leq m \leq M} |a^T_{c,k,m}|)]\). The results were obtained using a non-linear transforms learned with optimally choose values (by using cross-validation) of the parameter \( \lambda \) for the two different databases. As expected, we can see that the extreme points of the ratio between the \( \ell^*_\text{norms} \) \( C_1 \) and \( C_2 \) of the similarity concentrations and the discrimination power is around the optimal values of the parameter \( \tau = \lambda_1 \).

REFERENCES


