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

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper proposes learning a non-linear transform with two priors. The first is a discriminative prior defined using a measure on a support intersection and the second is a minimum information loss prior expressed as a constraint on the conditioning and the coherence. An approximation of the measure for the discriminative prior is addressed, connecting it to a similarity concentration. Along quantifying the discriminative properties of the transform representation a sensitivity analysis of the similarity concentration w.r.t. the parameters of the non-linear transform is given. Furthermore, a measure, related to the similarity concentration, reflecting the discriminative properties, named as discriminative power is introduced and its bounds are presented.

To support and validate the theoretical analysis a learning algorithm with the proposed prior is presented. The advantages and the potential of the proposed algorithm are evaluated by a computer simulation.

1 INTRODUCTION

Learning a transform that provides a sparse and discriminative representation is an active domain of research in many areas, including 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 representation 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), con, Jiang et al. (2016), Vu & Monga (2016b) address the estimate of the dictionary in a supervised meaner 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 $\mathbf{x}_{c,k} \in \Re^N$ as an approximation by a linear combination $\mathbf{y}_{c,k} \in \Re^M$ (refereed to as a sparse data representation) of a few words (vectors) $\|\mathbf{y}_{c,k}\|_0 << M$, from a dictionary (vector set) $\mathbf{D} \in \Re^{N \times M}$, *i.e.*, $\mathbf{x}_{c,k} = \mathbf{D}\mathbf{y}_{c,k} + \mathbf{v}_{c,k}$, $\mathbf{v}_{c,k} \in \Re^N$, with $\mathbf{v}_{c,k}$ 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), Vu & Monga (2016b) 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 same/different classes. The discrimination is enforced by approximating the prior with a structural constraints on the dictionary or imposing a discriminative term on the sparse representation. Additionally, in a number of works such as Mairal et al. (2008), Guo et al. (2012), Taalimi et al. (2015) is considered even a joint estimation/learning of dictionary, sparse representation and classifier by using iterative alternating minimization strategy. A comprehensive overview covering different approaches is given in Bengio et al. (2012), Cai et al. (2014) and Gangeh et al. (2015).

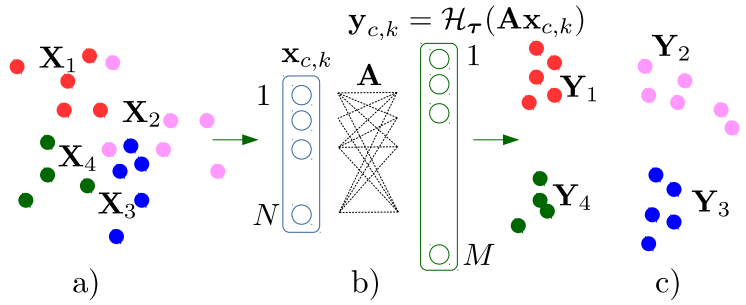


Figure 1: a) Data samples $\mathbf{X}_c, c \in \{1, 2, 3, 4\}$ from four different classes, b) Given a k-th data sample $\mathbf{x}_{c,k}$ from class c, the non-linear transform is represented as two step operation: linear mapping $\mathbf{A}\mathbf{x}_{c,k}$ (*step 1*) followed by an element wise thresholding function $\mathbf{y}_{c,k} = \mathcal{H}_{\boldsymbol{\tau}}(\mathbf{A}\mathbf{x}_{c,k})$ (*step 2*). c) The transform data samples $\mathbf{Y}_c, c \in \{1, 2, 3, 4\}$.

1.1 Open Issues

The general open issue for the 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) and Vu & Monga (2016b) 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.

In therms of the specifics about 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. Vu & Monga (2016c) proposed extension considering a low rank constraints on the dictionary. The authors in Guo et al. (2013) used a synthesis model with a pairwise constraints on the sparse representation. They have modeled the pair-wise constraints using a ℓ_2 distance metric. The methods of Yang et al. (2011b), Vu & Monga (2016c) and Guo et al. (2013) take into account assumption on the metric by defining the scatter and the pair-wise relation. Therefore, they constrain the space of the representation, that essentially is determined by the dictionary. However, these works do not consider whether that the used metric is optimal for the sparse representation. The authors in Liu et al. (2016) have proposed a method that 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 overlapped signal support of different classes, denoted as $c1$ and $c2$ is minimized. Furthermore, assuming $\mathbf{y}_{c1,k1} \in \Re^M$ and $\mathbf{y}_{c2,k2} \in \Re^M$ are two sparse representations for data samples $\mathbf{x}_{c1,k1} \in \Re^N$ and $\mathbf{x}_{c2,k2} \in \Re^N$, then the proposed similarity measure in Liu et al. (2016) is defined as expectation of $\|\mathbf{y}_{c1,k1} \odot \mathbf{y}_{c2,k2}\|_0$, where $\odot$ represents the Hadamard product. Note that two transform data samples $\mathbf{y}_{c1,k2}$ and $\mathbf{y}_{c2,k2}$ that have small support overlap $\|\mathbf{y}_{c1,k1} \odot \mathbf{y}_{c2,k2}\|_0 = s, s << M$, might not necessarily be similar or dissimilar, *i.e.*, $\mathbf{y}_{c1,k1} = \mathbf{y}_{c2,k2}$ and $\mathbf{y}_{c1,k1} = -\mathbf{y}_{c2,k2}$ with $\|\mathbf{y}_{c1,k1}\|_0 = \|\mathbf{y}_{c2,k2}\|_0 = s$ and s small.

1.2 Motivation and Approach

This paper addresses a *non-linear transform model*, where the data reconstruction is not modeled, and at least it is not targeted explicitly. Moreover, the model does not restrict the transform representation to be in the column space of the dictionary, allowing a rich class of representations to be modeled. In this line, we have to mention that a special case of the *non-linear transform model* is the *sparsifying transform model*, first introduced in Ravishankar & Bresler (2012). This model assumes that the data sample $\mathbf{x}$ is approximately sparsifiable under a linear transform $\mathbf{A} \in \Re^{M \times N}$, *i.e.*, $\mathbf{A}\mathbf{x}_{c,k} = \mathbf{y}_{c,k} + \mathbf{z}_{c,k}$, $\mathbf{z}_{c,k} \in \Re^M$, where $\mathbf{y}_{c,k}$ is sparse $\|\mathbf{y}_{c,k}\|_0 << M$. Moreover, we note that

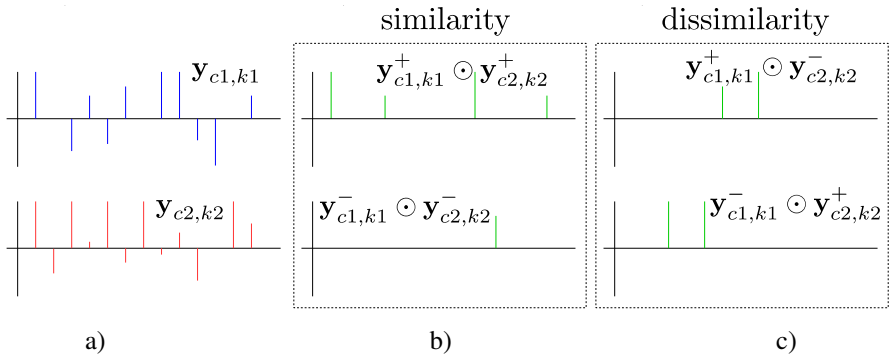


Figure 2: a) Two transform representations $\mathbf{y}_{c1,k1}$ and $\mathbf{y}_{c1,k1}$, b) the resulting Hadamard products $\mathbf{y}^+_{c,k} \odot \mathbf{y}^+_{c2,k2}$ and $\mathbf{y}^-_{c1,k1} \odot \mathbf{y}^-_{c2,k2}$ on the support intersection for the similarity contribution and c) the resulting Hadamard products $\mathbf{y}^+_{c1,k1} \odot \mathbf{y}^-_{c2,k2}$ and $\mathbf{y}^-_{c1,k1} \odot \mathbf{y}^+_{c2,k2}$ on the support intersection for the dissimilarity contribution between $\mathbf{y}c1, k1$ and $\mathbf{y}_{c2,k2}$.

the sparsifiying transform model represents a generalization of the analysis model Rubinstein et al. (2010), Rubinstein et al. (2013), Ravishankar & Bresler (2014) and Rubinstein & Elad (2014).

Together with the non-linear transform model, a novel prior is proposed, which is defined by a measure on the *support intersection* for the positive and negative components of a representation. The first motivation behind this prior is that that the support intersection on the positive and negative components between a transform data allows more freedom in imposing a regularization on the discriminative properties for the transform representation without taking into account any additional assumptions, contrary to Yang et al. (2011b), Vu & Monga (2016c), Guo et al. (2013) and Liu et al. (2016). Second, by a simple approximation on the prior the focus of the regularization is directly put on the contributing components for similarity/dissimilarity, *i.e.*, consider the measure $(\mathbf{y}^+_{c,k1})^T\mathbf{y}^+_{c,k2} + (\mathbf{y}^-_{c,k1})^T\mathbf{y}^-_{c,k2}$[1] between two transform representations $\mathbf{y}_{c,k1}$ and $\mathbf{y}_{c,k2}$. This measure captures the main contributing components for similarity on the support intersection, whereas $(\mathbf{y}^+_{c,k1})^T\mathbf{y}^-_{c,k2} + (\mathbf{y}^-_{c,k1})^T\mathbf{y}^+_{c,k2}$ captures the main contribution for dissimilarity between $\mathbf{y}_{c,k1}$ and $\mathbf{y}_{c,k2}$. Third, the expectation of this similarity measure across a the data set captures the *concentration of similarity* for that set. Therefore, it provides the possibility to define a formal notion to quantify the discriminative properties. In addition, this similarity measure is not ambiguous as the measure proposed in Liu et al. (2016) w.r.t. a notion for similarity/dissimilarity between two sparse representations $\mathbf{y}_{c1,k1}$ and $\mathbf{y}_{c2,k2}$. This is because the support intersections for the positive and the negative components $\|\mathbf{y}^+_{c1,k1} \odot \mathbf{y}^+_{c2,k2}\|_1$ and $\|\mathbf{y}^-_{c1,k1} \odot \mathbf{y}^-_{c2,k2}\|_1$, respectively, are considered separately.

A schematic diagram of the transform and the main idea behind the proposed concept are shown in Figure 1. In Figure 2 a), b), c) is given an illustrative example for the support intersection between the positive and negative component of two vectors $\mathbf{y}_{c1,k1}$ and $\mathbf{y}_{c2,k2}$ in the transform domain.

We propose a learning algorithm for the *non-linear transform model* with the discriminative prior. In a direction of quantifying the discriminative properties we give a sensitivity analysis w.r.t. the parameters of the transform. Along the same line, we present an information preservation relations for the change of a similarity concentrations w.r.t. change of the parameters of the transform,

[1] $\mathbf{y}_{c,k1} = \mathbf{y}^+_{c,k1} - \mathbf{y}^-_{c,k1}$ where $\mathbf{y}^+_{c,k1} = \max(\mathbf{y}_{c,k1}, \mathbf{0})$ and $\mathbf{y}^-_{c,k1} = \max(-\mathbf{y}_{c,k1}, \mathbf{0})$.

Note that $(\mathbf{y}^+_{c,k1})^T\mathbf{y}^+_{c,k2} + (\mathbf{y}^-_{c,k1})^T\mathbf{y}^-_{c,k2} = \|\mathbf{y}^+_{c,k1} \odot \mathbf{y}^+_{c,k2}\|_1 + \|\mathbf{y}^-_{c,k1} \odot \mathbf{y}^-_{c,k2}\|_1$ captures the only contribution for the similarity (whereas $\|\mathbf{y}^+_{c,k1} \odot \mathbf{y}^-_{c,k2}\|_1 + \|\mathbf{y}^-_{c,k1} \odot \mathbf{y}^+_{c,k2}\|_1$ captures the only contribution for the dissimilarity) between the vectors $\mathbf{y}_{c,k1}$ and $\mathbf{y}_{c,k2}$.

Moreover, it can be shown that for certain models under some conditions $\mathbf{y}^T_{c,k1}\mathbf{y}_{c,k2} = \|\mathbf{y}_{c,k1} \odot \mathbf{y}_{c,k2}\|_1$, if and only if the dissimilarity contribution on the support intersection between the vectors $\mathbf{y}_{c,k}$ and $\mathbf{y}_{c1,k1}$ is 0. That is $\mathbf{y}^T_{c,k2}\mathbf{y}_{c,k2} = (\mathbf{y}^+_{c,k1})^T\mathbf{y}^+_{c,k2} + (\mathbf{y}^-_{c,k1})^T\mathbf{y}^-_{c,k2} - (\mathbf{y}^-_{c,k1})^T\mathbf{y}^+_{c,k2} - (\mathbf{y}^+_{c,k1})^T\mathbf{y}^-_{c,k2} \leq (\mathbf{y}^+_{c,k1})^T\mathbf{y}^+_{c,k2} + (\mathbf{y}^-_{c,k1})^T\mathbf{y}^-_{c,k2}$, now if the dissimilarity contribution $-(\mathbf{y}^-_{c,k})^T\mathbf{y}^+_{c1,k1} - (\mathbf{y}^+_{c,k})^T\mathbf{y}^-_{c1,k1}$ is zero then $\|\mathbf{y}_{c,k1} \odot \mathbf{y}_{c,k2}\|_1 = (\mathbf{y}^+_{c,k1})^T\mathbf{y}^+_{c,k2} + (\mathbf{y}^-_{c,k1})^T\mathbf{y}^-_{c,k2} = \mathbf{y}^T_{c,k}\mathbf{y}_{c1,k1}$.

$x(n), A(n,m)$	scalar
$\mathbf{x}_{c,k}, \mathbf{y}_{c,k}$	vector
$\mathbf{X}, \mathbf{X}_c, \mathbf{X}_{\setminus\{k\in c\}}, \mathbf{Y}, \mathbf{A}$	matrix
$\odot$	Hadamard product
$\|.\|_p$	ℓ_p-norm,
$\mathcal{P}_c$=$\{\mathbf{A} \in \Re^{M\times N}, \boldsymbol{\tau}_c \in \Re^M\}$	set of parameters
$\mathcal{T}^{\mathcal{P}}(.)$	non-linear parametric function with parameters $\mathcal{P}$
$D^{\mathcal{P}}_{\ell_1}(\mathbf{x}_{c1,k1}; \mathbf{x}_{c2,k2})$	similarity measure between $\mathcal{T}^{\mathcal{P}}(\mathbf{x}_{c1,k1})$ and $\mathcal{T}^{\mathcal{P}}(\mathbf{x}_{c1,k2})$
$\frac{\partial D^{\mathcal{P}}_{\ell_1}(\mathbf{x}_{c1,k1};\mathbf{x}_{c2,k2})}{\partial \mathbf{A}}$	first order derivative of $D^{\mathcal{P}}_{\ell_1}(\mathbf{x}_{c1,k1}; \mathbf{x}_{c2,k2})$ w.r.t. $\mathbf{A}$
$D^{\mathcal{P}}_{\ell_1}(\mathbf{X})$	concentration of similarity
$\frac{\partial D^{\mathcal{P}}_{\ell_1}(\mathbf{X})}{\partial \mathbf{A}}$	change of concentration w.r.t. change of $\mathbf{A}$
$\frac{\partial D^{\mathcal{P}}_{\ell_1}(\mathbf{X})}{\partial \mathbf{A}_o}\|_{\mathbf{A}_o=\mathbf{B}}$	change of concentration w.r.t. change of $\mathbf{A}_o$ evaluated at $\mathbf{B}$
$\mathcal{I}^o, \mathcal{I}^t$	discriminative power in original and transform domain

Table 1: a) Most important notations used thought the paper

connecting all together the proposed prior, the model and the used non-linear transform. Furthermore, w.r.t. the similarity concentrations a notion defined as *discriminative power*, reflecting the discriminate properties of the representation for a data set is introduced. Based on the sensitivity analysis we present lower and upper bound on the discriminative power in the transform domain. On the practical side, the advantages and the potential of the proposed algorithm are demonstrated by a numerical experiments using the Extended YALE B Georghiades et al. (2001), AR Martínez & Benavente (1998), Norb LeCun et al. (2004), Coil-20 Nene et al. (1996), Clatech101 LeCun et al. (2008), UKB Nistér & Stewénius (2006) and MNIST Lecun & Cortes data sets.

1.3 NOTATIONS

A scalar variable is denoted using the usual symbols, *i.e.*, x, a vector is denoted by a bold, low caps symbols, *i.e.*, $\mathbf{x}$, a matrix by bold, upper cap symbol, *i.e.*, $\mathbf{A}$. 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 ℓ_p-norm is denoted as $\|.\|_p$ and the nuclear norm as $\|.\|_*$. The $\odot$ symbol represents the Hadamard product. Throughout the paper it is assumed that a set of data samples $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_C] \in \Re^{N\times L}, L = CK$ from C classes is given and that every class $c \in \mathcal{C} = \{1, 2, 3, ..., C\}$ has K samples, $\mathbf{X}_c = [\mathbf{x}_{c,1}, \mathbf{x}_{c,2}, ..., \mathbf{x}_{c,K}] \in \Re^{N\times K}$, $\mathbf{x}_{c,k} \in \Re^N$, $\forall c \in \mathcal{C}, \forall k \in \mathcal{K} = \{1, 2, ..., K\}$. We denote a the transform data as $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_C] \in \Re^{M\times L}$, where $\mathbf{Y}_c = [\mathbf{y}_{c,1}, \mathbf{y}_{c,2}, ..., \mathbf{y}_{c,K}] \in \Re^{M\times K}$ and $\mathbf{y}_{c,k} \in \Re^M$. We denote $\mathbf{X}_{\setminus\{k\in c\}} = \left[\mathbf{X}_1, \mathbf{X}_2, ..\mathbf{X}_{c,\setminus k}, ...\mathbf{X}_C\right] \in \Re^{N\times(L-1)}$ as the matrix that has all the columns of $\mathbf{X}$, except the column $\mathbf{x}_{c,k} \in \Re^N$, where $\mathbf{X}_{c,\setminus k} = [\mathbf{x}_{c,1}, \mathbf{x}_{c,2}, ..., \mathbf{x}_{c,k-1}, \mathbf{x}_{c,k+1}, ..., \mathbf{x}_{c,K}] \in \Re^{N\times(K-1)}$ 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 $\mathcal{M} = \{1, 2, ..., M\}$. We denote $\mathcal{D}^M$ as the set of all $M \times M$ diagonal matrices with non-negative diagonal elements.

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

Assume that training data $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_C] \in \Re^{N\times CK}$ are given, consisting of C classes. Per every class $c \in \mathcal{C}$ there are K training samples, *i.e.*, $\mathbf{X}_c = [\mathbf{x}_{c,1}, \mathbf{x}_{c,2}, ..., \mathbf{x}_{c,K}]$. Furthermore, we assume that per class $c \in \mathcal{C}$ there exist an unknown non-linear functions defined by set of parameters:

$$\mathcal{P}_c = \{\mathbf{A} \in \Re^{M\times N}, \boldsymbol{\tau}_c \in \Re^M\}, \forall c \in \mathcal{C}, \tag{1}$$

that separates the data samples from different classes in the transform domain. The linear map $\mathbf{A}$ is shared for all of them, but, the parameters $\boldsymbol{\tau}_c$ per different class c from the corresponding non-linear transforms $\mathcal{P}_c$ are different. Additionally, given any $\mathbf{x}_{c,k}$ from class c where $k \in \mathcal{K}$ it is assumed that the non-linear transform is expressible in two steps, consisting of a linear mapping (step 1) followed

by an element-wise non-linearity (step 2), as follows:

$$\mathbf{x}_{c,k} \xrightarrow[\text{step 1}]{\mathbf{A}} \mathbf{A}\mathbf{x}_{c,k} \xrightarrow[\text{step 2}]{\mathcal{H}_{\boldsymbol{\tau}_c}(.)} \mathbf{y}_{c,k}, \tag{2}$$

where $\mathcal{H}_{\boldsymbol{\tau}_c}(.) : \Re^M \rightarrow \Re^M$ is a non-linear thresholding function with parameters $\boldsymbol{\tau}_c$. It is important to mention that the thresholding is done with a different thresholding parameter per different transform dimension, *i.e.*, $\tau_c(m), \forall m \in \mathcal{M}$.

The goal of learning a non-linear transform (2) is to estimate only one model defined by a set of parameters

$$\mathcal{P} = \{\mathbf{A} \in \Re^{M \times N}, \boldsymbol{\tau} = \tau \mathbf{1} \in \Re_+^M\}, \tag{3}$$

that is common for all data samples that come form all classes by taking into account a set of priors. That is, we try to find as accurate as possible approximation (3) to the models represented by a set of parameters $\mathcal{P}_c = \{\mathbf{A} \in \Re^{M \times N}, \boldsymbol{\tau}_c \in \Re^M\}$ with as small as possible loss in the discriminative properties of the representation in the transform domain.

2.1 THE NON-LINEAR TRANSFORM MODEL WITH A DISCRIMINATIVE PRIOR

The compact description of the non-linear transform (2) by a *non-linear model* is expressed as follows:

$$\mathbf{A}\mathbf{x}_{c,k} = \mathbf{y}_{c,k} + \mathbf{z}_{c,k} \text{ where } \mathbf{y}_{c,k} \quad = \mathcal{T}^{\mathcal{P}_c}(\mathbf{x}_{c,k}), \tag{4}$$

where the function $\mathcal{T}^{\mathcal{P}_c}(.) : \Re^N \rightarrow \Re^M$ is a parametric non-linear function that gives $\mathbf{y}_{c,k}$, by using the set of parameters $\mathcal{P}_c$. The term $\mathbf{z}_{c,k} = \mathbf{A}\mathbf{x}_{c,k} - \mathbf{y}_{c,k}$ is the *non-linear transform error* vector that represents the deviation of the transform data $\mathbf{A}\mathbf{x}_{c,k}$ from the targeted transform representation $\mathbf{y}_{c,k} = \mathcal{T}^{\mathcal{P}_c}(\mathbf{x}_{c,k})$ in the transform domain.

Assuming Gaussian distributed error vector, the prior on $\mathbf{z}_{c,k}$ is modeled as $p(\mathbf{x}_{c,k}|\mathbf{y}_{c,k}) \propto \exp(-\frac{\|\mathbf{A}\mathbf{x}_{c,k} - \mathbf{y}_{c,k}\|_2}{\beta_0})$, additionally, assuming that the non-linear function $\mathcal{T}^{\mathcal{P}_c}(\mathbf{x}_{c,k})$ gives sparse $\mathbf{y}_{c,k}$, then we have the improper prior on $\mathbf{y}_{c,k}$, defined as $p(\mathbf{y}_{c,k}) \propto \exp(-\frac{\|\mathbf{y}_{c,k}\|_1}{\beta_1})$.

This paper models the discriminative prior as $p(\boldsymbol{\tau}_c, |\mathbf{y}_{c,k}) \propto \exp(-\frac{D(\mathbf{y}_{c,k};\boldsymbol{\tau}_c)}{\beta_2})$ by using an parametric measure $D(\mathbf{y}_{c,k};\boldsymbol{\tau}_c)$ with parameters $\boldsymbol{\tau}_c$. Assuming that the measure $D(\mathbf{y}_{c,k};\boldsymbol{\tau}_c)$ is defined on the support intersection for the positive and the negative components of $\mathbf{y}_{c,k}^+, \boldsymbol{\tau}_c^+$ and $\mathbf{y}_{c,k}^-, \boldsymbol{\tau}_c^-$, respectively, we propose the following formulation:

$$D(\mathbf{y}_{c,k};\boldsymbol{\tau}_c) = \|\mathbf{y}_{c,k}^+ \odot \boldsymbol{\tau}_c^+\|_1 + \|\mathbf{y}_{c,k}^- \odot \boldsymbol{\tau}_c^-\|_1 \tag{5}$$

where we highlight that the true $p(\boldsymbol{\tau}_c)$, $\boldsymbol{\tau}_c$, are not known and the only prior about $D(\mathbf{y}_{c,k};\boldsymbol{\tau}_c)$ and $p(\boldsymbol{\tau}_c)$ is that they should be informative enough w.r.t. the discriminative properties of the transform representation $\mathbf{y}_{c,k}$. Furthermore, instead of estimating them explicitly, an approximation to $D(\mathbf{y}_{c,k};\boldsymbol{\tau}_c)$ is considered, based only on the concentrations of similarity on the support intersection for the positive and negative components of the transform data in the transform domain.

The proposed approximation captures the *expected similarity on the support intersection* for the positive and negative component between $\mathbf{y}_{c,k}$ and the set of transform representations $\mathbf{Y}_{\setminus\{k \in c\}}$ that come from all classes $c1$ different from c, *i.e.*, $c \neq c1$, and is defined as follows:

$$D(\mathbf{y}_{c,k};\boldsymbol{\tau}_c) \sim D_{\ell_1}^{\mathcal{P}}(\mathbf{X}) = \sum_{\substack{c1 \\ c1 \neq c}} \sum_{k1} \|\mathbf{y}_{c,k}^+ \odot \mathbf{y}_{c1,k1}^+\|_1 + \sum_{\substack{c1 \\ c1 \neq c}} \sum_{k1} \|\mathbf{y}_{c,k}^- \odot \mathbf{y}_{c1,k1}^-\|_1, \tag{6}$$

where the m-th element $y_{c,k}^+(m)$ of $\mathbf{y}_{c,k}^+$ is defined as $y_{c,k}^+(m) = \max(y_{c,k}(m), 0)$ and similarly, $y_{c,k}^-(m) = \max(-y_{c,k}(m), 0), \forall m \in \mathcal{M}$. The approximation (6) might also be seen as finite sample estimates of the the unknown variables $\boldsymbol{\tau}_c$[2].

[2]Note that since $D_{\ell_1}^{\mathcal{P}}(\mathbf{X}) = \sum_{\substack{c1 \\ c1 \neq c}} \sum_{k1} \|\mathbf{y}_{c,k}^+ \odot \mathbf{y}_{c1,k1}^+\|_1 + \sum_{\substack{c1 \\ c1 \neq c}} \sum_{k1} \|\mathbf{y}_{c,k}^- \odot \mathbf{y}_{c1,k1}^-\|_1 = \| \left(\sum_{\substack{c1 \\ c1 \neq c}} \sum_{k1} \mathbf{y}_{c1,k1}^+ \right) \odot \mathbf{y}_{c,k}^+\|_1 + \| \left(\sum_{\substack{c1 \\ c1 \neq c}} \sum_{k1} \mathbf{y}_{c1,k1}^- \right) \odot \mathbf{y}_{c,k}^-\|_1$, $\boldsymbol{\tau}_c^- \sim \sum_{\substack{c1 \\ c1 \neq c}} \sum_{k1} \mathbf{y}_{c1,k1}^-$ and $\boldsymbol{\tau}_c^+ \sim \sum_{\substack{c1 \\ c1 \neq c}} \sum_{k1} \mathbf{y}_{c1,k1}^+$

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, all of them are with specific assumptions on the distribution of the data representation in the transform domain. The advantage of using (6) is that the aproximation is without any prior on the probability distributions $p(\boldsymbol{\tau}_c)$ and without any explicit assumption about the distance metric/measure, or the space/manifold in the transform domain.

2.2 THE LEARNING ALGORITHM

The summary of the used priors is as follows:

$$\begin{aligned} p(\mathbf{x}_{c,k}|\mathbf{y}_{c,k}) &\propto \exp(-\frac{\|\mathbf{A}\mathbf{x}_{c,k}-\mathbf{y}_{c,k}\|_2}{\beta_0}) \\ p(\mathbf{y}_{c,k}) &\propto \exp(-\frac{\|\mathbf{y}_{c,k}\|_1}{\beta_1}) \\ p(\boldsymbol{\tau}_c,|\mathbf{y}_{c,k}) &\propto \exp(-\frac{D(\mathbf{y}_{c,k};\boldsymbol{\tau}_c)}{\beta_2}). \end{aligned} \tag{7}$$

Additionally, we have a prior on $\mathbf{A}$ that penalizes the information loss in order to avoid trivially unwanted matrices $\mathbf{A}$, *i.e.*, matrices that have repeated or zero rows. The corresponding prior is defined as

$$p(\mathbf{A}) \propto \exp\left(-\left(\frac{1}{\beta_3}\|\mathbf{A}\|_F^2+\frac{1}{\beta_4}\|\mathbf{A}\mathbf{A}^T-\mathbf{I}\|_F^2-\frac{1}{\beta_5}\log|\det\mathbf{A}^T\mathbf{A}|\right)\right), \tag{8}$$

where the $\|\mathbf{A}\|_F$ penalty helps regularize the scale ambiguity, the $\log|\det(\mathbf{A}^T\mathbf{A})|$ and $\|\mathbf{A}\|_F^2$ are functions of the singular values of $\mathbf{A}$ and together help regularize the conditioning of $\mathbf{A}$. Assuming that the expected coherence $\mu^2(\mathbf{A})$ between the rows $\mathbf{a}_m$ of $\mathbf{A}$ (*i.e.*, $\mathbf{A}^T=[\mathbf{a}_1,\mathbf{a}_2,...,\mathbf{a}_M]$) is defined as $\mu^2(\mathbf{A})=\frac{2}{M(M-1)}\sum_{m_1\neq m_2}|\mathbf{a}_{m_1}\mathbf{a}_{m_2}^T|^2, \forall m_1,m_2\in\{1,2,..,M\}$. Then $\|\mathbf{A}\mathbf{A}^T-\mathbf{I}\|_F^2$ measures the expected coherence $\mu^2(\mathbf{A})$ and the ℓ_2 norm for the rows of $\mathbf{A}$.

Note that the joint probability can be expressed as:

$$\begin{aligned} p(\mathbf{x}_{c,k},\mathbf{y}_{c,k},\boldsymbol{\tau}_c,\mathbf{A}) = \quad & p(\mathbf{x}_{c,k},\mathbf{y}_{c,k},\boldsymbol{\tau}_c,|\mathbf{A})p(\mathbf{A}) = \\ & p(\mathbf{A}|\mathbf{x}_{c,k},\mathbf{y}_{c,k},\boldsymbol{\tau}_c)p(\mathbf{x}_{c,k},\mathbf{y}_{c,k},\boldsymbol{\tau}_c,) \\ p(\mathbf{x}_{c,k},\mathbf{y}_{c,k},\boldsymbol{\tau}_c) = \quad & p(\mathbf{x}_{c,k}|\mathbf{y}_{c,k})p(\boldsymbol{\tau}_c,|\mathbf{y}_{c,k})p(\mathbf{y}_{c,k}), \end{aligned} \tag{9}$$

since $p(\mathbf{x}_{c,k}|\mathbf{y}_{c,k},\boldsymbol{\tau}_c,)=p(\mathbf{x}_{c,k}|\mathbf{y}_{c,k})$.

Given the avlable tranining data set $\mathbf{X}\in\Re^{N\times CK}$ maximizing $p(\mathbf{x}_{c,k},\mathbf{y}_{c,k},\boldsymbol{\tau}_c,\mathbf{A})$ over $\mathbf{Y}$ and $\mathbf{A}$ is same as minimizing the following problem:

$$\begin{aligned} \underset{\mathbf{Y},\mathbf{A}}{min}\frac{1}{2}\|\mathbf{A}\mathbf{X}-\mathbf{Y}\|_F^2+\sum_{c,k}\lambda_0 D(\mathbf{y}_{c,k};\boldsymbol{\tau}_c)+\lambda_1\|\mathbf{y}_{c,k}\|_1+ \\ \frac{\lambda_2}{2}\|\mathbf{A}\|_F^2+\frac{\lambda_3}{2}\|\mathbf{A}\mathbf{A}^T-\mathbf{I}\|_F^2-\lambda_4\log|\det\mathbf{A}^T\mathbf{A}|, \end{aligned} \tag{10}$$

where $p(\mathbf{x}_{c,k},\mathbf{y}_{c,k},\boldsymbol{\tau}_c,\mathbf{A})$ is iteratively, marginally maximized over $\mathbf{Y}$ and $\mathbf{A}$ which is equivalent to maximizing the conditional densities $p(\mathbf{y}_{c,k}|\mathbf{x}_{c,k},\boldsymbol{\tau}_c,\mathbf{A})$ and $p(\mathbf{A}|\mathbf{y}_{c,k},\mathbf{x}_{c,k})$, respectively. In the following text we describe the iterating steps that consist of *linear map estimation* and *discriminative encoding*.

Linear map estimation: Given the available data samples $\mathbf{X}$ and the corresponding transform representations $\mathbf{Y}$ the linear map $\mathbf{A}$ estimation problem reduces to:

$$\underset{\mathbf{A}}{min}\frac{1}{2}\|\mathbf{A}\mathbf{X}-\mathbf{Y}\|_F^2+\frac{\lambda_2}{2}\|\mathbf{A}\|_F^2+\frac{\lambda_3}{2}\|\mathbf{A}\mathbf{A}^T-\mathbf{I}\|_F^2-\lambda_4\log|\det\mathbf{A}^T\mathbf{A}|, \tag{11}$$

and we use the ϵ-close closed form solution as proposed in Kostadinov et al. (2017) (*we give the proof in Appendix A*).

Algorithm 1 The learning algorithm

Input $\mathbf{X}, \lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4$
$\mathbf{A} \leftarrow inicialize$
repeat
  DISCRIMINATIVE ENCODING closed form solution per data sample
    $\mathbf{Y} \leftarrow \mathbf{AX}$
    repeat
      for $\forall c \in \mathcal{C}$ **do**
        $\mathbf{d}_c^- \leftarrow \sum_{\substack{c1 \\ c1 \neq c}} \sum_{k1} \mathbf{y}_{c1,k1}^-$ and $\mathbf{d}_c^+ \leftarrow \sum_{\substack{c1 \\ c1 \neq c}} \sum_{k1} \mathbf{y}_{c1,k1}^+$
      end for
      for $\forall c \in \mathcal{C}$ and $\forall k \in \mathcal{K}$ **do**
        $\mathbf{g} \leftarrow \text{sign}(\max(\mathbf{A}\mathbf{x}_{c,k}, \mathbf{0})) \odot \mathbf{d}_c^+ + \text{sign}(\max(-\mathbf{A}\mathbf{x}_{c,k}, \mathbf{0})) \odot \mathbf{d}_c^-$
        $\mathbf{y}_{c,k} \leftarrow \max\left(\mathbf{A}\mathbf{x}_{c,k} - \lambda_0 \mathbf{g} + \lambda_1 \mathbf{1}, \mathbf{0}\right) - \max\left(-\mathbf{A}\mathbf{x}_{c,k} - \lambda_0 \mathbf{g} + \lambda_1 \mathbf{1}, \mathbf{0}\right)$
        $\mathbf{y}_{c,k} \leftarrow \mathbf{y}_{c,k} / \|\mathbf{y}_{c,k}\|_2$
      end for
    until $convergence$
  TRANSFORM UPDATE ϵ-close closed form solution
    $\mathbf{U}_X \mathbf{\Sigma}_X^2 \mathbf{U}_X^T \leftarrow \mathbf{X}\mathbf{X}^T + \lambda_2 \mathbf{I}$ and $\mathbf{U}_{U_X XY} \mathbf{\Sigma}_{U_X XY} \mathbf{V}_{U_X XY}^T \leftarrow \mathbf{U}_X^T \mathbf{X} \mathbf{Y}^T$
      $\min_{\sigma_A(n)} \frac{\lambda_3}{{\sigma_X}^4} \sigma_A^4(n) + \left(\frac{\sigma_X^2(n) - 2\lambda_3}{\sigma_X^2(n)}\right) \sigma_A^2(n) - \frac{\sigma_\Gamma(n)}{\sigma_X(n)} \sigma_A(n) - 2\lambda_4 \log \frac{\sigma_A(n)}{\sigma_X(n)}$
        where $\sigma_\Gamma(n) \leftarrow T(n,n)$, $\mathbf{T} \leftarrow \mathbf{U}_{U_X XY} \mathbf{\Sigma}_{U_X XY}\ \mathbf{U}_{U_X XY}^T, \forall n \in \mathcal{N}$
    $\mathbf{A} \leftarrow \mathbf{V}_{U_X XY} \mathbf{U}_{U_X XY}^T \mathbf{\Sigma}_A \mathbf{\Sigma}_X^{-1} \mathbf{U}_X^T$
until $convergence$
Output $\mathbf{A}, \mathbf{Y}$

Discriminative encoding: Given the available data samples $\mathbf{X}$ and the current estimate of the transform $\mathbf{A}$, define for simplicity $\mathbf{Q} = \mathbf{AX} \in \Re^{M \times L}$. The discriminative representation estimation problem is formulated as:

$$\underset{\mathbf{Y}}{min} \|\mathbf{Q} - \mathbf{Y}\|_F^2 + \sum_{c,k} \lambda_0 D(\mathbf{y}_{c,k}; \boldsymbol{\tau}_c) + \lambda_1 \|\mathbf{y}_{c,k}\|. \tag{12}$$

Note that (10) and consequently (12) is a well defined problem only if the model variables $\boldsymbol{\tau}_c$, or the probability distributions $p(\boldsymbol{\tau}_c)$ are know in advance. Nevertheless, we show that the approximation (6) leads to an efficient solution of (12). Assuming that $\mathbf{Y}_{\setminus\{k \in c\}}$ is given, then for any sample $k \in \mathcal{K}$ from any class $c \in \mathcal{C}$ and the approximation (6), problem (12) represents a *constrained projection* taking the form of:

$$\underset{\mathbf{y}_{c,k}}{min} \|\mathbf{A}\mathbf{x}_{c,k} - \mathbf{y}_{c,k}\|_2^2 + \lambda_0 D(\mathbf{y}_{c,k}; \boldsymbol{\tau}_c) + \lambda_1 \|\mathbf{y}_{c,k}\|_1, \tag{13}$$

and has a closed form solution as:

$$\mathbf{y}_{c,k} = \max\left(\mathbf{A}\mathbf{x}_{c,k} - \lambda_0 \mathbf{g} + \lambda_1 \mathbf{1}, \mathbf{0}\right) - \max\left(-\mathbf{A}\mathbf{x}_{c,k} - \lambda_0 \mathbf{g} + \lambda_1 \mathbf{1}, \mathbf{0}\right), \tag{14}$$

where $\mathbf{g} = \text{sign}(\max(\mathbf{A}\mathbf{x}_{c,k}, \mathbf{0})) \odot {\mathbf{d}_c}^+ + \text{sign}(\max(-\mathbf{A}\mathbf{x}_{c,k}, \mathbf{0})) \odot {\mathbf{d}_c}^-$, $\mathbf{d}_c^- = \sum_{\substack{c1 \\ c1 \neq c}} \sum_{k1} \mathbf{y}_{c1,k1}^-$ and $\mathbf{d}_c^+ = \sum_{\substack{c1 \\ c1 \neq c}} \sum_{k1} \mathbf{y}_{c1,k1}^+$ (*the proof is given in Appendix B*).

We note that at convergence (which we do not prove here) we can only claim that a joint local maximum in $(\mathbf{Y}, \mathbf{A})$ of $p(\mathbf{x}_{c,k}, \mathbf{y}_{c,k}, \boldsymbol{\tau}_c, \mathbf{A})$ has been reached, even if, as in this case, each optimization step achieves the (marginal) ϵ-close and global optimal solution, respectively.

3 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.

$$\begin{array}{ccc} \text{original domain} & & \text{transform domain} \\ \mathbf{X} & \xrightarrow{\mathcal{T}^{\mathcal{P}}} & \mathbf{Y} \\ \updownarrow \mathcal{T}^{\mathcal{B}^N} & & \updownarrow \mathcal{T}^{\mathcal{B}^M} \\ \mathbf{X} & & \mathbf{Y} \\ \text{transform domain} & & \text{transform domain} \end{array}$$

Figure 3: The original and the transform domains under a non-linear transforms with a set of parameters $\mathcal{B}^N$, $\mathcal{P}$ and $\mathcal{B}^M$, note that for $\mathcal{T}^{\mathcal{B}^N}$ and $\mathcal{T}^{\mathcal{B}^M}$ the original and the transform domains are the same.

3.1 Sensitivity analysis w.r.t. the similarity concentration measures

To measure the ability for an increase in discriminative properties by a non-linear transform[3] 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 model* and the proposed non-linear transform model (3).

Any data set $\mathbf{X}$ in the original domain might have a transform model with parameters $\mathcal{B}^N = \{\mathbf{A}_o \in \Re^{N\times N}, \boldsymbol{\tau} = \mathbf{0} \in \Re_+^N\}$, if $\mathbf{A}_o = \mathbf{I} \in \mathcal{D}_+^N$ we refer to it as a *base original model*. Similarly as in the original domain, any data set $\mathbf{Y}$ in the transform domain might have a transform model with parameters $\mathcal{B}^M = \{\mathbf{A}_t \in \Re^{M\times M}, \boldsymbol{\tau} = \mathbf{0} \in \Re_+^M\}$, if $\mathbf{A}_t = \mathbf{I} \in \mathcal{D}_+^M$ we refer to it as a *base transform model*. Any base model, defined ether in the original domain $\mathcal{B}^N$ or in the transform domain $\mathcal{B}^M$, has domain equal to the co-domain, since $\mathbf{x}_{c,k} = \mathcal{T}^{\mathcal{B}^N}(\mathbf{x}_{c,k})$ and $\mathbf{y}_{c,k} = \mathcal{T}^{\mathcal{B}^M}(\mathbf{y}_{c,k})$ holds trivially, for the respective sets of parameters $\mathcal{B}^N = \{\mathbf{A}_o = \mathbf{I} \in \mathcal{D}_+^N, \boldsymbol{\tau} = \mathbf{0} \in \Re_+^N\}$ and $\mathcal{B}^M = \{\mathbf{A}_t = \mathbf{I} \in \mathcal{D}_+^M, \boldsymbol{\tau} = \mathbf{0} \in \Re_+^M\}$. This is illustrated with a diagram shown in Figure 3.

A base original model provides a possibility to compare it with any other non-linear transform model with parameters $\mathcal{P} = \{\mathbf{A} \in \Re^{M\times N}, \boldsymbol{\tau} \in \Re_+^M\}$. Additionally, note that for $\mathcal{B}^M = \{\mathbf{A}_t = \mathbf{I} \in \mathcal{D}^M, \boldsymbol{\tau} = \mathbf{0} \in \Re^M\}$ we have that $D_{\ell_1,c}^{\mathcal{B}^M}(\mathbf{Y}) = D_{\ell_1,c}^{\mathcal{P}}(\mathbf{X})$ and that $D_{\ell_1}^{\mathcal{B}^M}(\mathbf{Y}) = D_{\ell_1}^{\mathcal{P}}(\mathbf{X})$. It implies that the similarity concentrations can be analyzed as a function in the original domain under model $\mathcal{P}$ or in the transform domain under model $\mathcal{B}^M$. 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 Proposition 1.

Lemma 1: *The non-linear transform model* (4) *totally preserves the information in the change for the similarity concentration for a data set* $\mathbf{X}$ *w.r.t. a small change in the parameters of the models* $\mathcal{B}^N$, $\mathcal{B}^M$ *and* $\mathcal{P}$ *if* $\|\boldsymbol{\delta}_o\|_* = 0$ *and* $\|\boldsymbol{\delta}_t\|_* = 0$ *as*

$$\begin{aligned} (R1): \mathbf{A}\left(\frac{\partial D_{\ell_1,c}^{\mathcal{B}^N}(\mathbf{X})}{\partial \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}} + \frac{\partial D_{\ell_1}^{\mathcal{B}^N}(\mathbf{X})}{\partial \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}\right) &= \frac{\partial D_{\ell_1,c}^{\mathcal{P}}(\mathbf{X})}{\partial \mathbf{A}} + \frac{\partial D_{\ell_1}^{\mathcal{P}}(\mathbf{X})}{\partial \mathbf{A}} + \boldsymbol{\delta}_o \\ (R2): \mathbf{A}\left(\frac{\partial D_{\ell_1,c}^{\mathcal{P}}(\mathbf{X})}{\partial \mathbf{A}} + \frac{\partial D_{\ell_1}^{\mathcal{P}}(\mathbf{X})}{\partial \mathbf{A}}\right)^T &= \frac{\partial D_{\ell_1,c}^{\mathcal{B}^M}(\mathbf{Y})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} + \frac{\partial D_{\ell_1}^{\mathcal{B}^M}(\mathbf{Y})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} + \boldsymbol{\delta}_t \\ (R3): \mathbf{A}\boldsymbol{\delta}_o^T &= \frac{\partial D_{\ell_1,c}^{\mathcal{B}^M}(\mathbf{Z})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} + \frac{\partial D_{\ell_1}^{\mathcal{B}^M}(\mathbf{Z})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} + \boldsymbol{\delta}_t^T, \end{aligned} \tag{15}$$

where $\boldsymbol{\delta}_o = \sum_{c,c1}\sum_{\substack{k,k1\\k\neq k1}} \mathbf{z}_{c1,k1}\mathbf{x}_{c,k}^T + \mathbf{z}_{c,k}\mathbf{x}_{c1,k1}^T$, $\frac{\partial D_{\ell_1,c}^{\mathcal{P}}(\mathbf{X})}{\partial \mathbf{A}} + \frac{\partial D_{\ell_1}^{\mathcal{P}}(\mathbf{X})}{\partial \mathbf{A}} = \sum_{c,c1}\sum_{\substack{k,k1\\k\neq k1}} \mathbf{y}_{c1,k1}\mathbf{x}_{c,k}^T + \mathbf{y}_{c,k}\mathbf{x}_{c1,k1}^T$, $\boldsymbol{\delta}_t = \sum_{c,c1}\sum_{\substack{k,k1\\k\neq k1}} \mathbf{z}_{c1,k1}\mathbf{y}_{c,k}^T + \mathbf{z}_{c,k}\mathbf{y}_{c1,k1}^T$, $\frac{\partial D_{\ell_1,c}^{\mathcal{B}^M}(\mathbf{Y})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} + \frac{\partial D_{\ell_1}^{\mathcal{B}^M}(\mathbf{Y})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} = \sum_{c,c1}\sum_{\substack{k,k1\\k\neq k1}} \mathbf{y}_{c1,k1}\mathbf{y}_{c,k}^T + \mathbf{y}_{c,k}\mathbf{y}_{c1,k1}^T$, $\frac{\partial D_{\ell_1,c}^{\mathcal{B}^M}(\mathbf{Z})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} + \frac{\partial D_{\ell_1}^{\mathcal{B}^M}(\mathbf{Z})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} =$

[3] instead of using the term non-liner transform model as defined by (2) or (4) for short we just use the term model

$\sum_{c,c1}\sum_{\substack{k,k1\\k\neq k1}}\mathbf{z}_{c1,k1}\mathbf{z}_{c,k}^T+\mathbf{z}_{c,k}\mathbf{z}_{c1,k1}^T$, $\frac{\partial D_{\ell_1,c}^{\mathcal{B}^N}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}=\sum_c\sum_{k,k1}\mathbf{x}_{c,k}\mathbf{x}_{c,k1}^T+\mathbf{x}_{c,k1}\mathbf{x}_{c,k}^T$ *and* $\frac{\partial D_{\ell_1}^{\mathcal{B}^N}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}=\sum_{c,c1,c\neq c1}\sum_{k,k1}\mathbf{x}_{c,k}\mathbf{x}_{c1,k1}^T+\mathbf{x}_{c1,k1}\mathbf{x}_{c,k}^T$. *The proof is given in Appendix C.*

The terms $\mathbf{z}_{c,k}$ represent the non-linear transform error vectors that appear in the model $\mathbf{A}\mathbf{x}_{c,k}=\mathbf{y}_{c,k}+\mathbf{z}_{c,k}$ as a result of applying an element-wise non-liner operation $\mathcal{H}_{\boldsymbol{\tau}}$ to $\mathbf{A}\mathbf{x}_{c,k}$, *i.e.*, $\mathbf{y}_{c,k}=\mathcal{H}_{\boldsymbol{\tau}}(\mathbf{A}\mathbf{x}_{c,k})$. As an example in the sparsifing transform model $\mathbf{z}_{c,k}$ is the "loss of information", that is the information about the values of the elements in $\mathbf{A}\mathbf{x}_{c,k}$ that are discarded. The terms $\boldsymbol{\delta}_o$ and $\boldsymbol{\delta}_t$ correlate the errors $\mathbf{z}_{c,k}$ with the original data $\mathbf{x}_{c1,k1}$ and transform data $\mathbf{y}_{c1,k1}$, 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 $\boldsymbol{\delta}_o=\mathbf{0}$ and $\boldsymbol{\delta}_t=\mathbf{0}$. Moreover, $\boldsymbol{\delta}_o$ and $\boldsymbol{\delta}_t$ bear important information about the discriminative properties in the transform domain.

The terms $\frac{\partial D_{\ell_1,c}^{\mathcal{P}}(\mathbf{X})}{\partial\mathbf{A}}$ and $\frac{\partial D_{\ell_1}^{\mathcal{P}}(\mathbf{X})}{\partial\mathbf{A}}$ represent the change of the similarity concentrations under infinitesimally small change of the parameter $\mathbf{A}$ from the model $\mathcal{P}$. The terms $\frac{\partial D_{\ell_1,c}^{\mathcal{B}^N}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}$ and $\frac{\partial D_{\ell_1}^{\mathcal{B}^N}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}$ have dual interpretation. Assuming metric $\mathbf{A}_o=\mathbf{I}$, 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.* $\mathbf{x}_{c,k}\sim\mathcal{N}(\boldsymbol{\mu}=\mathbf{0},\boldsymbol{\Sigma}=\mathbf{I})$, then $\frac{\partial D_{\ell_1,c}^{\mathcal{B}^N}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}$ and $\frac{\partial D_{\ell_1}^{\mathcal{B}^N}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}$ 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 $\mathcal{B}^N$ with any arbitrary transform defined in the original domain $\mathcal{P}$. The relation $(R2)$ is a result about the preservation of change in the similarity concentration between two models $\mathcal{B}^M$ and $\mathcal{P}$ 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 relation between: *the linear projection* (*by the linear map* $\mathbf{A}$ *that appears in the model* $\mathcal{P}$) *of the change in the similarity concentration under the model* $\mathcal{B}^N$ *in the original domain and the change of the similarity concentration under the model* $\mathcal{B}^M$ *in the transform domain*.

This relation exists independently for $\frac{\partial D_{\ell_1,c}^{\mathcal{B}^N}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}$ and $\frac{\partial D_{\ell_1}^{\mathcal{B}^N}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}$, nevertheless, we will define the summarized and the independent versions. Therefore, first we define $\frac{\partial\mathcal{J}_{\ell_1}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}=\frac{\partial D_{\ell_1,c}^{\mathcal{B}^N}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}+\frac{\partial D_{\ell_1}^{\mathcal{B}^N}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}$ and $\frac{\partial\mathcal{J}_{\ell_1}(\mathbf{Y})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}=\frac{\partial D_{\ell_1,c}^{\mathcal{B}^N}(\mathbf{Y})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}+\frac{\partial D_{\ell_1}^{\mathcal{B}^N}(\mathbf{Y})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}$ and rewrite $(R1)$ as $\mathbf{A}\frac{\partial\mathcal{J}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}-\boldsymbol{\delta}_o=\frac{\partial D_{\ell_1,c}^{\mathcal{P}}(\mathbf{X})}{\partial\mathbf{A}}+\frac{\partial D_{\ell_1}^{\mathcal{P}}(\mathbf{X})}{\partial\mathbf{A}}$, then replace $\frac{\partial D_{\ell_1,c}^{\mathcal{P}}(\mathbf{X})}{\partial\mathbf{A}}+\frac{\partial D_{\ell_1}^{\mathcal{P}}(\mathbf{X})}{\partial\mathbf{A}}$ in $(R2)$ by the same term in $(R1)$, use $(R3)$, reorder and we have the following result.

Lemma 2: *For fixed* τ *any non-linear transform model* (4) *preservers the information in the change of similarity concentrations w.r.t. a small change in* $\mathbf{A}$ *by*

$$
\begin{aligned}
(R4):&\mathbf{A}\frac{\partial\mathcal{J}_{\ell_1}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}\mathbf{A}^T=\frac{\partial\mathcal{J}_{\ell_1}(\mathbf{V})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}=\frac{\partial\mathcal{J}_{\ell_1}(\mathbf{Y})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}+\boldsymbol{\xi}_c+\boldsymbol{\xi},\\
(R5):&\mathbf{A}\frac{\partial D_{\ell_1,c}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}\mathbf{A}^T=\frac{\partial D_{\ell_1,c}(\mathbf{V})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}=\frac{\partial D_{\ell_1,c}(\mathbf{Y})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}+\boldsymbol{\xi}_c,\\
(R6):&\mathbf{A}\frac{\partial D_{\ell_1}(\mathbf{X})}{\partial\mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}\mathbf{A}^T=\frac{\partial D_{\ell_1}(\mathbf{V})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}=\frac{\partial D_{\ell_1}(\mathbf{Y})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}+\boldsymbol{\xi},
\end{aligned}
\tag{16}
$$

where $\mathbf{V}=\mathbf{A}\mathbf{X}$, $\boldsymbol{\xi}_c+\boldsymbol{\xi}=\frac{\partial\mathcal{J}_{\ell_1}(\mathbf{Z})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}+\frac{\partial\mathcal{J}_{\ell_1}(\mathbf{Y};\mathbf{Z})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}$, $\frac{\partial\mathcal{J}_{\ell_1}(\mathbf{Z})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}=\frac{\partial D_{\ell_1,c}^{\mathcal{B}^N}(\mathbf{Z})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}+\frac{\partial D_{\ell_1}^{\mathcal{B}^N}(\mathbf{Z})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}$ *and* $\frac{\partial\mathcal{J}_{\ell_1}(\mathbf{Y};\mathbf{Z})}{\partial\mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}=\boldsymbol{\delta}_t+\boldsymbol{\delta}_t^T$.

$$\begin{array}{ccc} \mathbf{X} & \xrightarrow{\mathcal{T}^{\mathcal{P}}} & \mathbf{Y} \\ \downarrow & & \downarrow \\ \log \frac{D^{\mathcal{B}^N}_{\ell_1,c}(\mathbf{X})}{D^{\mathcal{B}^N}_{\ell_1}(\mathbf{X})+\epsilon} & & \log \frac{D^{\mathcal{P}}_{\ell_1,c}(\mathbf{X})}{D^{\mathcal{P}}_{\ell_1}(\mathbf{X})+\epsilon} = \log \frac{D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})}{D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})+\epsilon} \\ \downarrow & & \downarrow \\ \mathcal{I}^o & & \mathcal{I}^t \end{array}$$

Figure 4: The relation for the definition of the discriminative power in the original and the transform domain under the base models $\mathcal{B}^N$ and $\mathcal{B}^M$.

Note that $\frac{\partial^n \mathcal{J}_{\ell_1}(\mathbf{X})}{\partial^n \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}} = 4\frac{\partial^{n-1} \mathcal{J}_{\ell_1}(\mathbf{X})}{\partial^{n-1} \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}$ and $\frac{\partial^n \mathcal{J}_{\ell_1}(\mathbf{Y})}{\partial^n \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} = 4\frac{\partial^{n-1} \mathcal{J}_{\ell_1}(\mathbf{Y})}{\partial^{n-1} \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}$ therefore, $\frac{\partial \mathcal{J}_{\ell_1}(\mathbf{X})}{\partial \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}} = \frac{1}{4}\frac{\partial^2 \mathcal{J}_{\ell_1}(\mathbf{X})}{\partial^2 \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}$ and $\frac{\partial \mathcal{J}_{\ell_1}(\mathbf{Y})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} = \frac{1}{4}\frac{\partial^2 \mathcal{J}_{\ell_1}(\mathbf{Y})}{\partial^2 \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}$ might be interpreted as Fisher information matrices evaluated at $\mathbf{A}_o = \mathbf{I} \in \mathcal{D}^N$ and $\mathbf{A}_t = \mathbf{I} \in \mathcal{D}^M$. The expressions by $(R4)$ actually relate the metric in the original domain under the model $\mathcal{B}^N$ to the induced metric in the transform domain for the model $\mathcal{B}^M$, with induction done by the model $\mathcal{P}$ with parameter set $\{\mathbf{A} \in \Re^{M\times N}, \boldsymbol{\tau} \in \Re^M_+\}$. 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 $\mathcal{B}^N$ and $\mathcal{B}^M$ are smooth the analysis of their relations reveals insights about the relation between the manifolds under the models $\mathcal{B}^N$ and $\mathcal{P}$. The terms $\frac{\partial \mathcal{J}_{\ell_1}(\mathbf{Z})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}$ and $\frac{\partial \mathcal{J}_{\ell_1}(\mathbf{Y};\mathbf{Z})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}$ carry out the information about the breaks and the discontinuities of the regularity and smoothness in the manifold induced by the model $\mathcal{P}$. Furthermore, if $\|\frac{\partial \mathcal{J}_{\ell_1}(\mathbf{Z})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}\|_* = 0$ and $\|\frac{\partial \mathcal{J}_{\ell_1}(\mathbf{Y};\mathbf{Z})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}\|_* = 0$, then $(R4)$ in information geometry Amari (2013) is seen as change of coordinates on a manifold, where the intrinsic properties of curvature remain unchanged under different parametrization.

3.2 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 discriminative power, based on a measure for the relations between the concentrations $D^{\mathcal{P}}_{\ell_1,c}(\mathbf{X})$ and $D^{\mathcal{P}}_{\ell_1}(\mathbf{X})$.

Proposition 1: *The discriminative power for any dataset $\mathbf{X} \in \Re^{M\times CK}$ under any transform with parameter set $\mathcal{P}$ is defined as:*

$$\mathcal{I}^t = \log(D^{\mathcal{P}}_{\ell_1,c}(\mathbf{X})) - \log(D^{\mathcal{P}}_{\ell_1}(\mathbf{X}) + \epsilon) = \log(D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})) - \log(D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y}) + \epsilon). \tag{17}$$

Remark 1: *The advantage of this measure is that it logarithmically signifies the difference between $D^{\mathcal{P}}_{\ell_1,c}(\mathbf{X})$ and $D^{\mathcal{P}}_{\ell_1}(\mathbf{X})$*[4] .

The definition about the discriminative power of the data set $\mathbf{X}$, but, now under a model with a parameter set $\mathcal{B}^N$ is equivalent to the one defined for $\mathcal{I}^t$, we denote it as $\mathcal{I}^o$. An illustration is given by a diagram shown in Figure 4.

The bound on the discriminative power is given by the following result.

Theorem 1: *The discriminative power for any data set $\mathbf{X} \in \Re^{N\times CK}$ under any transform with parameter set $\mathcal{P}$ is bounded as:*

$$\log(\lambda_{min}(\mathbf{A}^T\mathbf{A})) + \log(\frac{Tr\{\frac{\partial D^{\mathcal{B}^N}_{\ell_1,c}(\mathbf{X})}{\partial \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}\}}{D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{AX}) + \epsilon}) \leq \mathcal{I}^t \leq \log(D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{AX})) - \log(\epsilon) \tag{18}$$

The proof is given in Appendix D.

[4]Another relation is the difference $D^{\mathcal{P}}_{\ell_1,c}(\mathbf{X}) - D^{\mathcal{P}}_{\ell_1}(\mathbf{X})$ between $D^{\mathcal{P}}_{\ell_1,c}(\mathbf{X})$ and $D^{\mathcal{P}}_{\ell_1}(\mathbf{X})$ that actually represents a finite sample approximation to a *discriminative density* since it approximates the density $\log(\frac{p(\boldsymbol{\tau}_c,\mathbf{y}_{c,k})}{p(\boldsymbol{\upsilon}_c,\mathbf{y}_{c,k})})$, *i.e.*, $D^{\mathcal{P}}_{\ell_1,c}(\mathbf{X}) - D^{\mathcal{P}}_{\ell_1}(\mathbf{X}) \sim \log(\frac{p(\boldsymbol{\tau}_c,\mathbf{y}_{c,k})}{p(\boldsymbol{\upsilon}_c,\mathbf{y}_{c,k})}) = D(\boldsymbol{\tau}_c,\mathbf{y}_{c,k}) - D(\boldsymbol{\upsilon}_c,\mathbf{y}_{c,k}) = \|\boldsymbol{\tau}_c{}^+ \odot \mathbf{y}^+_{c,k}\|_1 + \|\boldsymbol{\tau}_c{}^- \odot \mathbf{y}^-_{c,k}\|_1 - (\|\boldsymbol{\upsilon}_c{}^+ \odot \mathbf{y}^+_{c,k}\|_1 + \|\boldsymbol{\upsilon}_c{}^- \odot \mathbf{y}^-_{c,k}\|_1)$, where $\boldsymbol{\upsilon}_c \in \Re^M$ is an unknown parameter.

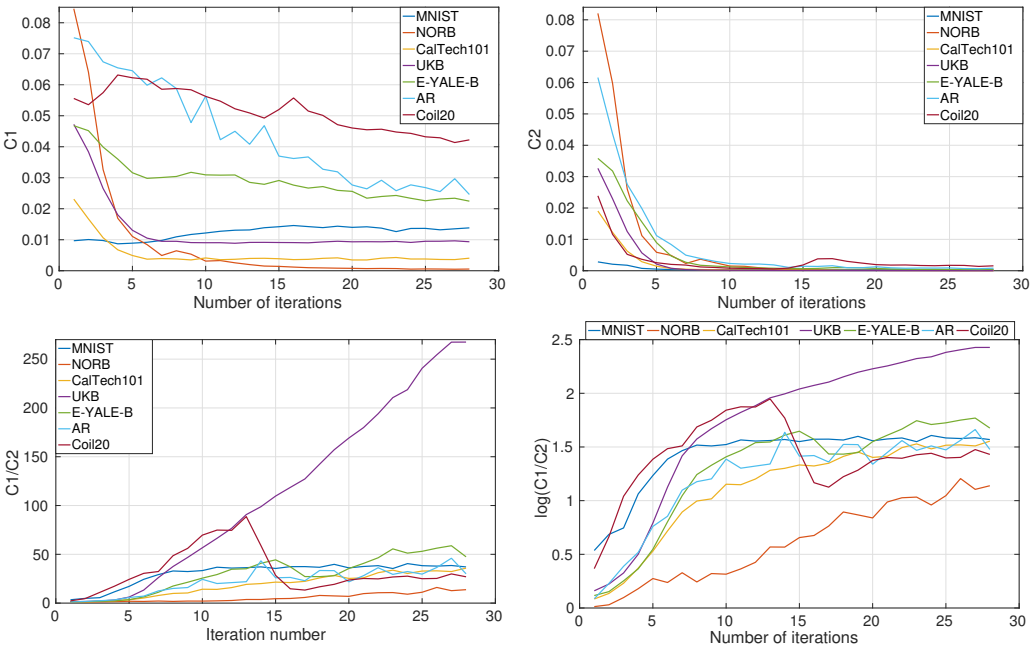


Figure 5: The evolution of the similarity concentrations $C1 = D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})$ and $C2 = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})$, their ratio $C1/C2$ and the discriminative power $\log(C1/C2) = \mathcal{I}^t$ during the learning of the non-linear transform with transform dimension $M = 19000$.

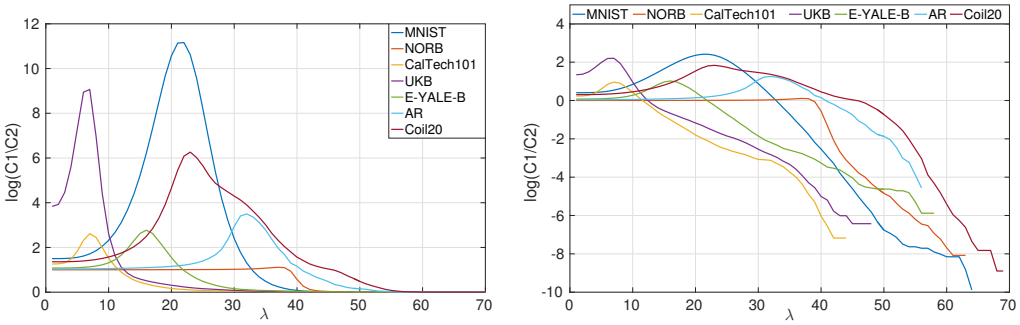


Figure 6: The ratio $C1/C2$ of the similarity concentrations $C1 = D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})$ and $C2 = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})$ and the discriminative power $\log(C1/C2) = \mathcal{I}^t$ for randomly chosen subsets from all of the used databases under a non-linear transform with transform dimension $M = 19000$ and varying thresholding parameter $\boldsymbol{\tau} = \lambda\mathbf{1}$.

At first the resulting bounds might look counterintuitive since the loss of information seems to increase the discriminative 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 $\mathcal{P}$ for which there is no loss of information $\|\boldsymbol{\xi}_c + \boldsymbol{\xi}\|_* = 0$, the only condition for the increase in the discriminative power is $D^{\mathcal{B}^{\mathcal{N}}}_{\ell_1}(\mathbf{X}) \geq D^{\mathcal{B}^M}_{\ell_1}(\mathbf{AX})$ and $D^{\mathcal{B}^{\mathcal{N}}}_{\ell_1,c}(\mathbf{X}) \leq D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{AX})$.

Second, in the rest of the cases for which $D^{\mathcal{B}^{\mathcal{N}}}_{\ell_1}(\mathbf{X}) \geq D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})$ and $D^{\mathcal{B}^{\mathcal{N}}}_{\ell_1,c}(\mathbf{X}) \leq D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})$ holds true it will be possible to increase the discriminative power. Moreover, there is a trade-off between the increase in discriminative power as a result of loss of information $\|\boldsymbol{\xi}_c + \boldsymbol{\xi}\|_* > 0$.

	$C_n(\mathbf{A})$	$\mu(\mathbf{A})$	$t_e[min]$	$\mathcal{I}^O$	$\mathcal{I}^{RT}$	$\mathcal{I}^{ST^*}$	$\mathcal{I}^{NT}$
$D1$	2.21	0.03	5.10	0.03	0.18	0.68	**1.98**
$D2$	1.80	0.02	5.45	0.02	0.10	1.30	**1.79**
$D3$	2.12	0.02	6.55	0.00	0.01	0.71	**1.61**
$D4$	0.08	0.02	8.92	0.08	0.61	0.89	**1.89**
$D5$	6.01	0.01	12.8	0.01	0.16	1.02	**2.12**
$D6$	33.1	0.02	30.1	0.06	0.53	1.36	**3.36**
$D7$	1.60	0.02	5.00	0.13	0.63	1.06	**1.96**

Table 2: The conditioning number $C_n(\mathbf{A})$ and the expected mutual coherence $\mu(\mathbf{A})$ for the learned transform $\mathbf{A}$, The execution time $t_e[min]$ in minutes of the proposed algorithm for 28 iterations at the transform domain dimensionality $M = 19000$.

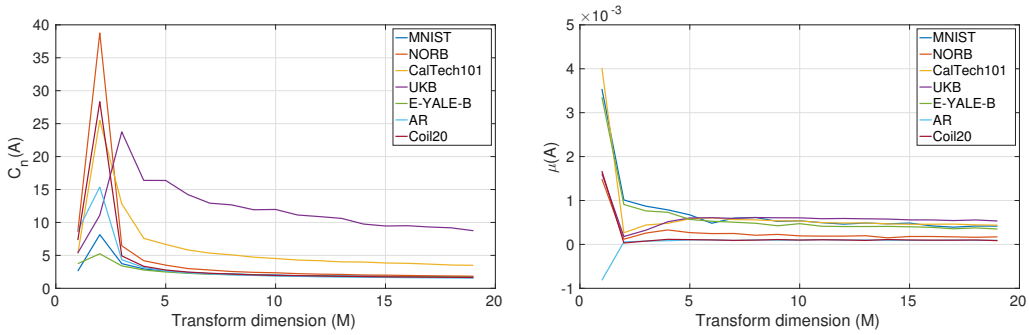


Figure 7: The conditioning number $C_n(\mathbf{A})$ and the expected mutual coherence $\mu(\mathbf{A})$ for the learned linear transform $\mathbf{A}$ at different dimensionality $M \in \{100, 1150, 2200, 3250, 4300, 5350, 6400, 7450, 8500, 9550, 10600, 11650, 12700, 13750, 14800, 15850, 16900, 17950, 19000\}$.

4 Numerical experiments

The numerical experiments are summarized in two different parts. In the first series of the experiments the properties of the learned map $\mathbf{A}$ for the proposed algorithm are investigated. We evaluate the computational efficiency, as run time $t_e[min]$, the conditioning number $C_n(\mathbf{A})$ [5], the expected mutual coherence $\mu(\mathbf{A})$ and the discriminative power across several databases for a learned non-linear transforms having different dimensionality. Additionally, a comparison between the resulting discriminative power 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 support dissimilarity prior, denoted as $\mathcal{I}^0$, $\mathcal{I}^{RT}$, $\mathcal{I}^{ST^*}$ and $\mathcal{I}^{NT}$, respectively, is estimated and presented .

The second part evaluates a comparison of the discriminative power between the proposed algorithm and different supervised dictionary learning methods (SDL) Ramirez et al. (2010), Yang et al. (2011a), Vu et al. (2015) and Vu & Monga (2016a). This comparison considers a setup where the used data sets are spited 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 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 ($D2$) Martínez & Benavente (1998), Norb ($D3$) LeCun et al. (2004), Coil-20 ($D4$) 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 downscaled to resolutions 21×21, 32×28, 24×24, 20×25, 21×21, 20×25, 28×28, respectively, and are normalized to unit variance. Considering the used implementation of the algorithm note that the singular value decomposition for large matrix has high computational complexity. However,

[5] the conditioning number is computed as the ratio between the largest and the smallest singular value of the matrix $\mathbf{A}$

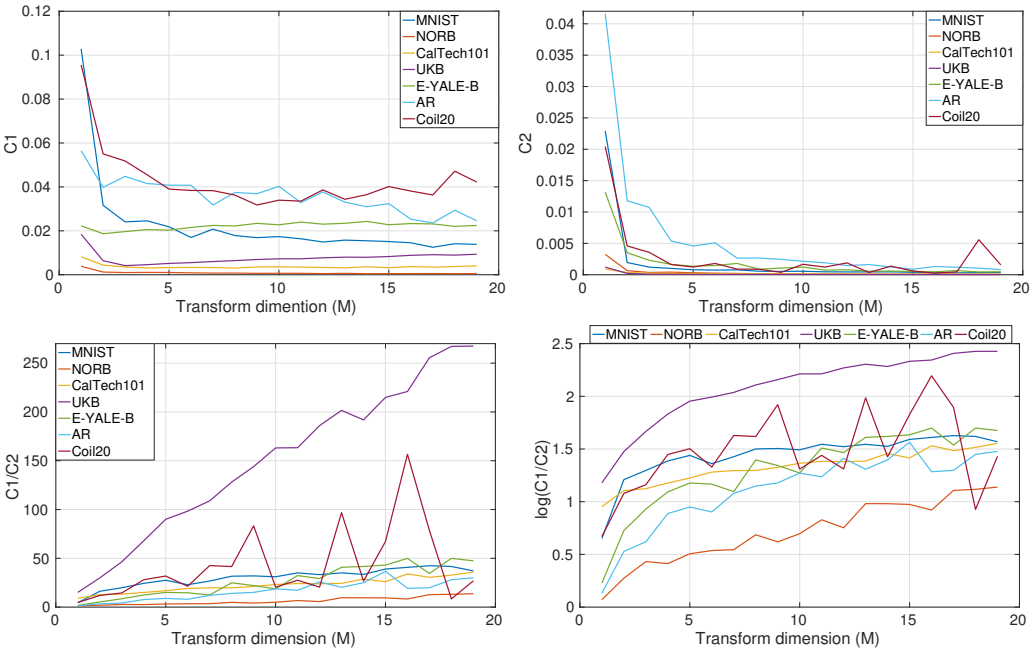


Figure 8: The similarity concentrations $C1 = D^{\mathcal{B}^M}_{\ell_{1,c}}(\mathbf{Y})$ and $C2 = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})$, their ratio $C1/C2$ and the discriminative power $\log(C1/C2) = \mathcal{I}^t$ on a subset of the transform data using learned non-linear transform at different dimensionality $M \in \{100, 1150, 2200, 3250, 4300, 5350, 6400, 7450, 8500, 9550, 10600, 11650, 12700, 13750, 14800, 15850, 16900, 17950, 19000\}$.

	$D1$	$D7$
$\mathcal{I}^{DLSI}$	0.71	0.67
$\mathcal{I}^{FDDL}$	0.87	0.63
$\mathcal{I}^{COPAR}$	0.57	0.54
$\mathcal{I}^{LRSDL}$	0.42	0.40
$\mathcal{I}^{NT}$	**0.98**	**0.81**

a)

	$D1$ Acc. [%]
$DLSI$	96.5
$FDDL$	97.5
$COPAR$	98.3
$LRSDL$	98.7
ST	**99.7**

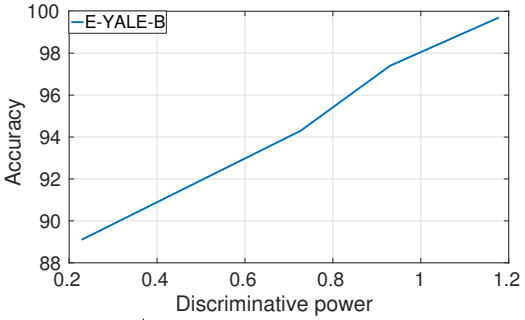
b)

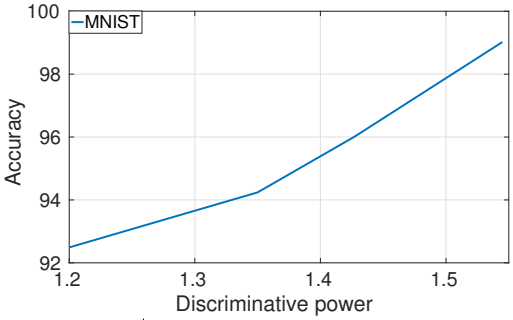
	$D7$ Acc. [%]
$DLSI$	98.74
$FDDL$	96.31
$COPAR$	96.41
$LRSDL$	–
ST	**99.02**

c)

Table 3: a) The discriminative power for the methods $DLSI$Ramirez et al. (2010), $FDDL$ Yang et al. (2011a), $COPAR$ Vu et al. (2015) and $LRSDL$ Vu & Monga (2016a) and the proposed sparsifying transform ST, b) and c) The recognition results on the Extended Yale B and MNIST database.

$\mathbf{A} - \hat{\mathbf{A}}$, 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 (11). 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 ρ 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). the parameters λ_0 and λ_1 are set such that the resulting non-linear transform representation has very small number of non-zeros w.r.t. the transform dimension, in the experiments here 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 average of 3 runs. An implementation presented in Vu & Monga (2016a) was used to learn the dictionaries and estimate the sparse codes for the respective supervised dictionary learning methods (SDL) Ramirez et al. (2010), Yang et al. (2011a), Vu et al. (2015) and Vu & Monga (2016a).





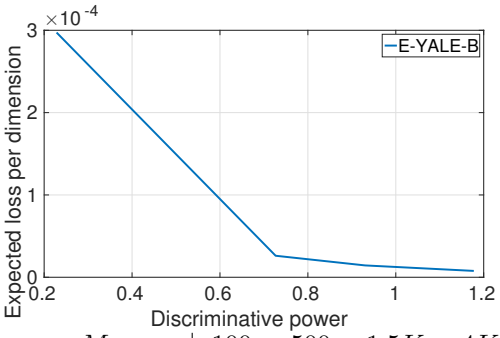
M	100	500	$1.5K$	$4K$
Acc. [%]	89.1	94.3	97.4	**99.7**
$\mathcal{I}^t$	0.21	0.73	0.93	1.23

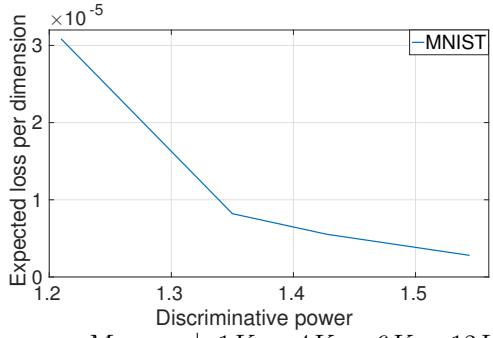
a)

M	$1K$	$4K$	$6K$	$12K$
Acc. [%]	92.49	94.24	96.01	**99.02**
$\mathcal{I}^t$	1.18	1.35	1.42	1.55

b)

Figure 9: a) and b) The recognition results and the discriminative 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.





M	100	500	$1.5K$	$4K$
$\mathbb{E}[\frac{\lVert \mathbf{z}_{c,k} \rVert_2}{M}]10^{-4}$	3	0.3	0.14	0.08
$\mathcal{I}^t$	0.21	0.73	0.93	1.23

a)

M	$1K$	$4K$	$6K$	$12K$
$\mathbb{E}[\frac{\lVert \mathbf{z}_{c,k} \rVert_2}{M}]10^{-5}$	3.1	0.82	0.6	0.28
$\mathcal{I}^t$	1.18	1.35	1.42	1.55

b)

Figure 10: a) and b) The expected loss per transform dimension $\mathbb{E}[\frac{\lVert \mathbf{z}_{c,k} \rVert_2}{M}] = \mathbb{E}[\lVert \mathbf{A}\mathbf{x}_{c,k} - \mathbf{y}_{c,k} \rVert_2 / M]$ and the discriminative power on the Extended Yale B and MNIST databases, respectively, on the transform representation $\mathbf{Y}$, obtained by using a non-linear transform $\mathcal{T}^{\mathcal{P}}$ at different dimensionality M.

Linear map properties, the change in the similarity concentrations and the discriminative power The conditioning number and the expected coherence for the learned transforms are shown on Table 2. 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 2. The learned transforms for all the data sets have relatively low execution time, regardless of a transform dimensions 19000. The discriminative power is significantly increased in the transform domain $\mathcal{I}^{NT}$ compared to the one in the original domain $\mathcal{I}^{O}$ and is higher than $\mathcal{I}^{ST^*}$ and $\mathcal{I}^{RT}$.

The evolution of the similarity concentrations $C1 = D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})$ and $C2 = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})$, their ratio $C1/C2$ and the discriminative power $\log(C1/C2) = \mathcal{I}^t$ for a subsets of the used databases after applying a non-linear transform with transform dimension $M = 19000$ is shown on Figure 5. It is important to note that the similarity concentrations $C1 = D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})$ and $C2 = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})$ are decreasing, but, the slope of decrease for $C2$ is stronger. Therefore, the discriminative 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.

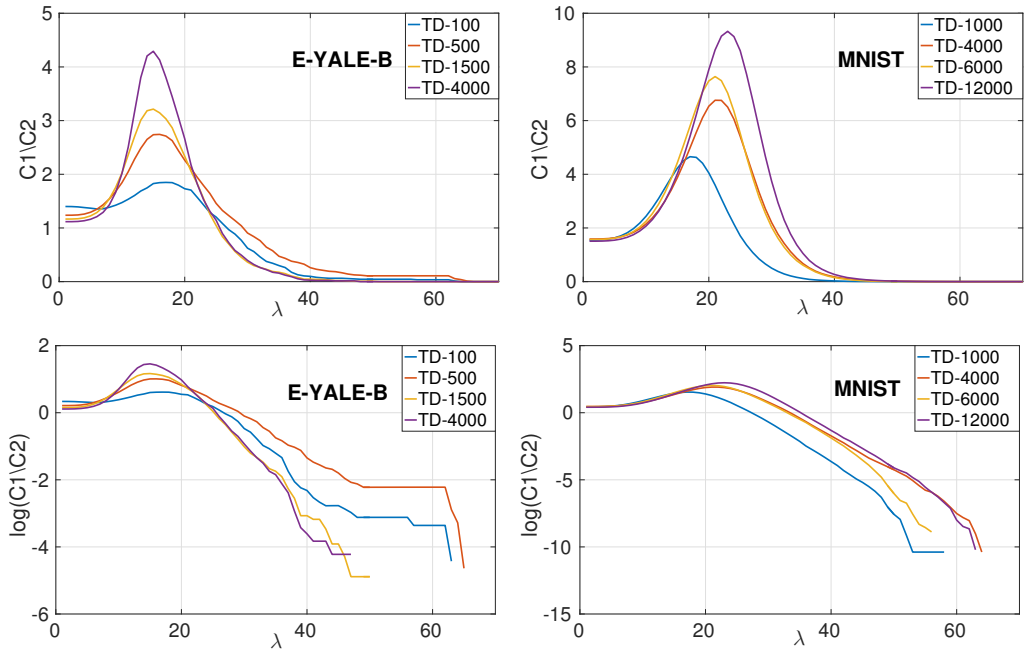


Figure 11: The ratio of the similarity concentrations similarity concentrations $C1 = D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})$ and $C2 = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})$ and the discriminative power $\log(C1/C2) = \mathcal{I}^t$ for the Extended Yale B and MNIST databases under non-linear transforms having different transform dimension M and varying thresholding parameter $\boldsymbol{\tau} = \lambda\mathbf{1}$.

The conditioning number and the expected coherence for the learned transforms for all the databases at different transform dimensions are shown on Figure 7. We see that the value of both the conditioning number and the coherence is reducing and converging to a common values, respectively, implying that the conditioning and the coherence constraints are effective.

The ratio $C1/C2$ between the similarity concentrations $C1 = D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})$ and $C2 = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})$ and the discriminative power $\log(C1/C2) = \mathcal{I}^t$ on a subsets of the used databases after applying a non-linear transform with transform dimension $M = 19000$ and varying the thresholding parameter $\boldsymbol{\tau} = \lambda\mathbf{1}$ is shown on Figure 6. We used 70 different values for the parameter λ, sampled uniformly from the interval $\big(0, (\max_{c,k} \max_m |\mathbf{a}_m^T \mathbf{x}_{c,k}|)\big)$. The result were obtained using a non-linear transforms learned with one value for the parameter λ 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 value of the parameter λ.

The similarity concentrations $C1 = D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})$ and $C2 = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})$, their ratio $C1/C2$ and the discriminative power $\log(C1/C2) = \mathcal{I}^t$ for a subsets of the used databases after applying a non-linear transform having transform dimensions $M \in \{100, 1150, 2200, 3250, 4300, 5350, 6400, 7450, 8500, 9550, 10600, 11650, 12700, 13750, 14800, 15850, 16900, 17950, 19000\}$ is shown on Figure 8. We can see a similar behavior as previous, that is, the similarity concentrations $C1 = D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})$ and $C2 = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})$ are decreasing, but, the slope of decrease for $C2 = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})$ is stronger. Therefore, the discriminative power increases as the transform dimension increases.

SNTL vs SDL discriminative 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 $LRSDL$ Vu & Monga (2016a). 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 evaluation. We compute both the discriminative power and the recognition accuracy on the test sets.

The dictionary size (transform dimension M) is set to be equal to $150, 75, 1515, 3825, 570, 150, 300$ for the corresponding databases, respectively, in all of the comparing algorithms. The discriminative power of the comparing methods is denoted as $\mathcal{I}^{DLSI}$, $\mathcal{I}^{FDDL}$, $\mathcal{I}^{COPAR}$ and $\mathcal{I}^{LRSDL}$, respectively. The recognition results for the methods $DLSI$, $FDDL$, $COPAR$ and $LRSDL$ on the data sets Extended YALE B and MNIST were not computed here, rather than that 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 (2016a). Considering the proposed algorithm the transform was learned for the transform dimensions M equal to $[100, 500, 1500, 4000]$ and $[1000, 4000, 6000, 12000]$, 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. Finally, the evaluation was performed on the respective test sets.

The results are shown on Table 3 a), b) and c). The discriminative power of the proposed non-linear transform is higher that the discriminative power of the comparing methods. Whereas the recognition accuracy is higher for high dimensionality of the proposed method and outperforms the SDL methods at dimensionality 4000 and 12000.

The results about the accuracy of recognition and the expected loss measured as $\mathbb{E}[\frac{\|\mathbf{z}_{c,k}\|_2}{M}] = \mathbb{E}[\|\mathbf{A}\mathbf{x}_{c,k} - \mathbf{y}_{c,k}\|_2/M]$ per the discriminative power at different transform dimension are shown on Figure 9 and Figure 10, respectively. It is interesting to highlight that as the discriminative power at different transform dimension increases it also increases the accuracy of recognition, moreover, the results on these two data bases show that this increase is approximately linear. On the other hand the expected loss per transform dimension decreases as the discriminative power at different transform dimension increases.

The results about the ratio $C1/C2$ between the similarity concentrations $C1 = D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})$ and $C2 = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})$ and the discriminative power $\log(C1/C2) = \mathcal{I}^t$ on Extended-Yale-B and MNIST databases after applying a non-linear transform with transform dimensions $M = \{100, 500, 1500, 4000\}$ and $M = \{1000, 4000, 6000, 12000\}$, respectively, and varying the thresholding parameter $\boldsymbol{\tau} = \lambda\mathbf{1}$ are shown on Figure 11. We again used 70 different values for the parameter λ, sampled uniformly from the interval $\big(0, (\max_{c,k} \max_m |\mathbf{a}_m^T \mathbf{x}_{c,k}|)\big)$, that were different for the used databases. The result were obtained using a non-linear transforms learned with optimally choose values (by using cross-validation) of the parameter λ for the two different databases. As expected we can see the extreme points of the ratio between the similarity concentrations $C1 = D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})$ and $C2 = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})$ and the discriminative power $\log(C1/C2) = \mathcal{I}^t$ is around the optimal values of the parameter $\boldsymbol{\tau} = \lambda\mathbf{1}$.

5 Conclusion

This paper presented an analysis on the discriminative properties for models expressible as two step transform, linear mapping (*step* 1) followed by an element wise non-linearity (*step* 2). A novel prior was proposed and the properties around the model and the proposed prior were investigated. A low complexity learning algorithm was presented with the proposed priors. The preliminary results w.r.t. the introduced measures and the recognition accuracy on the used databases showed promising performance.

A future analysis for the *synthesis model* and *transform based auto-encoder* towards generalization and fair comparison between different transforms and encoding methods is possible. In the line of the encoding one might consider to minimize directly the difference or the ratio between the introduced similarity concentrations or the discriminative power.

An extension covering the sufficient conditions for increase in discriminative power in the transform domain, together with an analysis for a deep architecture where per single layer we have one non-linear transform are left for our future work.

A.

THE GLOBAL OPTIMAL SOLUTION

Given the current estimate of $\mathbf{Y}$ the estimate of the transform $\mathbf{A}$ is solution to the following problem:

$$\{\hat{\mathbf{A}}\} = \underset{\mathbf{A}}{argmin} \, \|\mathbf{AX} - \mathbf{Y}\|_F^2 + \lambda_2 \|\mathbf{A}^T\mathbf{A}\|_F + \lambda_3 \|\mathbf{AA}^T - \mathbf{I}\|_F - \lambda_4 (\log|\det \mathbf{A}^T\mathbf{A}|). \tag{19}$$

Consider the equivalent trace form of (19) we have:

$$\begin{aligned} &\underset{\mathbf{A}}{min} Tr\{(\mathbf{AX} - \mathbf{Y})^T(\mathbf{AX} - \mathbf{Y})\} + \lambda_2 Tr\{\mathbf{A}^T\mathbf{A}\} + \lambda_3 Tr\{(\mathbf{AA}^T - \mathbf{I})^T(\mathbf{AA}^T - \mathbf{I})\} - \\ &\quad \lambda_4(\log|\det \mathbf{A}^T\mathbf{A}|). \end{aligned} \tag{20}$$

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

$$\mathbf{U}_X \mathbf{\Sigma}_X \mathbf{U}_X^T \mathbf{U}_X \mathbf{\Sigma}_X \mathbf{U}_X^T = \mathbf{XX}^T + \lambda_2 \mathbf{I}. \tag{21}$$

Let

$$\mathbf{A} = \mathbf{BD}, \mathbf{D} = \quad \mathbf{U}_X \mathbf{\Sigma}_X^{-1} \mathbf{U}_X^T. \tag{22}$$

Define

$$\begin{aligned} &g_1 = \mathbf{BDXY}^T, g_2 = \mathbf{BB}^T, g_3 = \left(\mathbf{BDD}^T\mathbf{B}\right)\left(\mathbf{BDD}^T\mathbf{B}\right)^T \\ &g_4 = g_5 = (\mathbf{BD})(\mathbf{BD})^T \sum_{m=1}^{M} \mathbf{P}_{m,M}(\mathbf{BD})(\mathbf{BD})^T \mathbf{P}_{m,M} \\ &g_6 = \sum_{m=1}^{M} \mathbf{P}_{m,M}(\mathbf{BD})(\mathbf{BD})^T \mathbf{P}_{m,M}, g_7 = \log|\det \mathbf{BDD}^T\mathbf{B}^T|. \end{aligned} \tag{23}$$

where $\mathbf{P}_{m,M} \in \mathcal{D}^M$, $P_{m,M}(m1, m1) = 1$ if $m1 = m$ and $P_{m,M}(m1, m1) = 0$ if $m1 \neq m$. Then (19) equivalently is:

$$\min_{\mathbf{B}} -Tr\{g_1\} + Tr\{g_2\} + \lambda_3(Tr\{g_3\} - Tr\{g_4\}) + \lambda_3\left(Tr\{g_5\} - Tr\{g_6\}\right) - \lambda_4 g_7. \tag{24}$$

Assume that $\mathbf{B}$ decomposes as:

$$\mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B^T, \tag{25}$$

where $\mathbf{\Sigma}_B$ is diagonal with positive diagonal elements, $\mathbf{U}_B$ is column orthogonal and $\mathbf{V}_B$ is orthogonal square matrix. Let

$$\mathbf{U}_B = \mathbf{U}_X \mathbf{V}_{XY}, \mathbf{V}_B = \quad \mathbf{U}_X^T, \tag{26}$$

then

$$Tr\{g_1\} = Tr\{\mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B \mathbf{U}_X \mathbf{\Sigma}_X^{-1} \mathbf{U}_X^T \mathbf{XY}^T\} = Tr\{\mathbf{\Sigma}_B \mathbf{\Sigma}_X^{-1} \mathbf{\Sigma}_{XY}\}. \tag{27}$$

The term

$$Tr\{g_2\} = Tr\{(\mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B)(\mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B)^T\} = Tr\{\mathbf{\Sigma}_B^2\}. \tag{28}$$

Define $\mathbf{\Sigma} = \mathbf{\Sigma}_B \mathbf{\Sigma}_X^{-2} \mathbf{\Sigma}_B$ then

$$Tr\{g_3\} = Tr\{\left(\mathbf{\Sigma}\mathbf{U}_B^T\mathbf{U}_B\right)\left(\mathbf{\Sigma}\mathbf{U}_B^T\mathbf{U}_B\right)^T\} = Tr\{\mathbf{\Sigma}\mathbf{\Sigma}\} = Tr\{\mathbf{\Sigma}_B^4 {\mathbf{\Sigma}_X}^{-4}\}. \tag{29}$$

Note that

$$\begin{aligned} &Tr\{g_4\} = \\ &Tr\{\mathbf{BDD}^T\mathbf{B}^T \sum_{m=1}^{M} \mathbf{P}_{m,M}\mathbf{BDD}^T\mathbf{B}^T\mathbf{P}_{m,M}\} = Tr\{\left(\sum_{m=1}^{M} \mathbf{P}_{m,M}\mathbf{U}_X\mathbf{\Sigma}\mathbf{U}_X^T\mathbf{P}_{m,M}\right)^2\}. \end{aligned} \tag{30}$$

Therefore, $\lambda_3 Tr\{g_4\} = \lambda_3 Tr\{g_5\}$. The term $Tr\{g_6\} = Tr\{\left(\mathbf{\Sigma}_B\mathbf{\Sigma}_X^{-1}\right)^2\}$. The term g_7 is

$$\log|\det \mathbf{A}^T\mathbf{A}| = \log|\det \mathbf{U}_X \mathbf{\Sigma}_Y^{-1} \mathbf{\Sigma}_B^2 \mathbf{\Sigma}_Y^{-1} \mathbf{U}_X| = |\det \mathbf{\Sigma}_Y^{-1} \mathbf{\Sigma}_B^2 \mathbf{\Sigma}_Y^{-1}|. \tag{31}$$

Finally (19) is reduced to:

$$\min_{\mathbf{\Sigma}_B} \sum_{n=1}^{N} \frac{\lambda_3}{\sigma_X^4(n)} \sigma_B^4(n) + \left(\frac{\sigma_X^2(n) - 2\lambda_3}{\sigma_X^2(i)} \right) \sigma_B^2(n) - \frac{\sigma_{U_X XY}(n)}{\sigma_X(n)} \sigma_B(n) - 2\lambda_4 \log \frac{\sigma_B(n)}{\sigma_X(n)}. \quad (32)$$

Equalling to zero the first order derivative w.r.t $\sigma_B(n)$ of the objective in (32) and multiplying by $\sigma_B(n)$ gives:

$$4 \frac{\lambda_3}{\sigma_X^4(i)} \sigma_B^4(i) + 2 \left(\frac{\sigma_X^2(i) - 2\lambda_3}{\sigma_X^2(i)} \right) \sigma_B^2(i) - \frac{\sigma_{U_X XY}(i)}{\sigma_X(i)} \sigma_B(i) - 2\lambda_4 = 0. \quad (33)$$

A closed form solution to (33) exists and depends on the discriminant Δ of the quartic polynomial. Moreover, since $4\frac{\lambda_4}{\sigma_X{}^4}$ is positive a global minimum exists □

THE ϵ-CLOSE SOLUTION

Consider the equivalent trace form of (19). Note that $\forall \lambda_2 \geq 0$, $\mathbf{X}\mathbf{X}^T + \lambda_2 \mathbf{I}$ is symmetric positive definite matrix with all eigenvalues non-negative, therefore it decomposes as

$$\mathbf{U}_X \mathbf{\Sigma}_X \mathbf{U}_X^T \mathbf{U}_X \mathbf{\Sigma}_X \mathbf{U}_X^T = \mathbf{X}\mathbf{X}^T + \lambda_2 \mathbf{I}. \quad (34)$$

Let

$$\mathbf{D} = \mathbf{U}_X \mathbf{\Sigma}_X^{-1} \mathbf{U}_X^T, \text{ and define } \mathbf{A} = \mathbf{B}\mathbf{D}, \mathbf{U}_{U_X XY} \mathbf{\Sigma}_{U_X XY} \mathbf{V}_{U_X XY}^T = \mathbf{U}_X \mathbf{X}\mathbf{Y}^T. \quad (35)$$

Assume that $\mathbf{B}$ decomposes as $\mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B^T$, where $\mathbf{\Sigma}_B$ is diagonal with positive diagonal elements, $\mathbf{U}_B$ is column orthogonal and $\mathbf{V}_B$ is orthogonal square matrix.

Let

$$\mathbf{U}_B = (\mathbf{U}_{U_X XY} \mathbf{V}_{U_X XY}^T)^T, \mathbf{V}_B = \mathbf{U}_X^T, \quad (36)$$

then

$$\begin{aligned} Tr\{\mathbf{A}\mathbf{X}\mathbf{Y}^T\} = Tr\{\mathbf{B}\mathbf{U}_X \mathbf{\Sigma}_X^{-1} \mathbf{U}_X^T \mathbf{X}\mathbf{Y}^T\} = Tr\{\mathbf{V}_{U_X XY} \mathbf{U}_{U_X XY}^T \mathbf{\Sigma}_B \\ \mathbf{\Sigma}_X^{-1} \mathbf{U}_{U_X XY} \mathbf{\Sigma}_{U_X XY} \mathbf{V}_{U_X XY}^T\} = Tr\{\mathbf{\Sigma}_\Gamma \mathbf{\Sigma}_B \mathbf{\Sigma}_X^{-1}\}. \end{aligned} \quad (37)$$

The term

$$Tr\{\mathbf{B}\mathbf{B}^T\} = Tr\{(\mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B)(\mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B)^T\} = Tr\{\mathbf{\Sigma}_B^2\}. \quad (38)$$

Define $\mathbf{\Sigma} = \mathbf{\Sigma}_B \mathbf{\Sigma}_X^{-2} \mathbf{\Sigma}_B$, then

$$Tr\{(\mathbf{A}\mathbf{A}^T)^2\} = Tr\{(\mathbf{\Sigma}\mathbf{U}_B^T \mathbf{U}_B)(\mathbf{\Sigma}\mathbf{U}_B^T \mathbf{U}_B)^T\} = Tr\{\mathbf{\Sigma}\mathbf{\Sigma}\} = Tr\{\mathbf{\Sigma}_B^4 \mathbf{\Sigma}_X{}^{-4}\}. \quad (39)$$

The term $Tr\{\mathbf{A}\mathbf{A}^T\} = Tr\{\left(\mathbf{\Sigma}_B \mathbf{\Sigma}_X^{-1}\right)^2\}$. The term $\log|\det \mathbf{A}^T\mathbf{A}|$ is

$$\log|\det \mathbf{A}^T\mathbf{A}| = \log|\det \mathbf{U}_X \mathbf{\Sigma}_Y^{-1} \mathbf{\Sigma}_B^2 \mathbf{\Sigma}_Y^{-1} \mathbf{U}_X| = |\det \mathbf{\Sigma}_Y^{-1} \mathbf{\Sigma}_B^2 \mathbf{\Sigma}_Y^{-1}|. \quad (40)$$

Finally (19) is reduced to

$$\min_{\mathbf{\Sigma}_B} \sum_{i=1}^{N} \frac{\lambda_3}{\sigma_X^4(n)} \sigma_B^4(n) + \left(\frac{\sigma_X^2(n) - 2\lambda_3}{\sigma_X^2(n)} \right) \sigma_B^2(n) - \frac{\sigma_{U_X XY}(n)}{\sigma_X(n)} \sigma_B(n) - 2\lambda_4 \log \frac{\sigma_B(n)}{\sigma_X(n)}. \quad (41)$$

Equaling to zero the first order derivative w.r.t $\sigma_B(n)$ of the objective for (41) and multiplying by $\sigma_B(n)$ gives:

$$4 \frac{\lambda_3}{\sigma_X^4(n)} \sigma_B^4(n) + 2 \left(\frac{\sigma_X^2(n) - 2\lambda_3}{\sigma_X^2(n)} \right) \sigma_B^2(n) - \frac{\sigma_{U_X XY}(n)}{\sigma_X(n)} \sigma_B(n) - 2\lambda_4 = 0. \quad (42)$$

A closed form solution to (41) exists and it depends on the discriminant Δ of the quartic polynomial. Moreover, since $4\frac{\lambda_4}{\sigma_X{}^4}$ is positive a global minimum to (41) exists.

Having $\mathbf{U}_B = (\mathbf{U}_{U_X XY}\ \mathbf{V}_{U_X XY}^T)^T$ and $\mathbf{V}_B = \mathbf{U}_X^T$ with the solution for $\mathbf{\Sigma}_B$ by (41) gives the ϵ-close solution to problem (19) as

$$\mathbf{A} = (\mathbf{U}_{U_X XY} \mathbf{V}_{U_X XY}^T)^T \mathbf{\Sigma}_B \mathbf{\Sigma}_X^{-1} \mathbf{U}_X, \quad (43)$$

where

$$\min_{\mathbf{\Sigma}_B} \max_{\mathbf{U}_B, \mathbf{V}_B} Tr\{\mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B^T \mathbf{U}_X \mathbf{\Sigma}_X^{-1} \mathbf{U}_X \mathbf{X}\mathbf{Y}^T\} \leq Tr\{\mathbf{\Sigma}_\Gamma \mathbf{\Sigma}_B \mathbf{\Sigma}_X^{-1}\} \quad (44)$$

implying that the ϵ-close closed form solution is a lower bound on the solution to (19).

APPENDIX B.

Let $\mathbf{y}_{c,k} = \mathbf{y}^+_{c,k} + \mathbf{y}^-_{c,k}, \mathbf{y}^+_{c,k} \in \Re^M_+$ and $\mathbf{y}^-_{c,k} \in \Re^M_-$. Note that the term $D^{\mathcal{P}}_{\ell_1}(\mathbf{X})$ is defined as:

$$\begin{aligned} D^{\mathcal{P}}_{\ell_1}(\mathbf{X}) = & \sum_{\substack{c1,c2 \\ c1\neq c2}} \sum_{k1,k2} \|\mathbf{y}^+_{c1,k1} \odot \mathbf{y}^+_{c2,k2}\|_1 + \sum_{\substack{c1,c2 \\ c1\neq c2}} \sum_{k1,k2} \|\mathbf{y}^-_{c1,k1} \odot \mathbf{y}^-_{c2,k2}\|_1 = \\ & \sum_{\substack{c1,c2 \\ c1\neq c2}} \sum_{k1,k2} |\mathbf{y}^+_{c1,k1}|^T |\mathbf{y}^+_{c2,k2}| + \sum_{\substack{c1,c2 \\ c1\neq c2}} \sum_{k1,k2} |\mathbf{y}^-_{c1,k1}|^T |\mathbf{y}^-_{c2,k2}| = \\ & |\mathbf{y}^+_{c1,k1}|^T \sum_{\substack{c1,c2 \\ c1\neq c2}} \sum_{k1,k2} |\mathbf{y}^+_{c2,k2}| + |\mathbf{y}^-_{c1,k1}|^T \sum_{\substack{c1,c2 \\ c1\neq c2}} \sum_{k1,k2} |\mathbf{y}^-_{c2,k2}| = \\ & |\mathbf{y}^+_{c1,k1}|^T \mathbf{g}^+ + |\mathbf{y}^-_{c1,k1}|^T \mathbf{g}^-, \end{aligned} \tag{45}$$

where $\mathbf{g}^+ = \sum_{\substack{c1,c2 \\ c1\neq c2}} \sum_{k1,k2} |\mathbf{y}^+_{c2,k2}|$, $\mathbf{g}^- = \sum_{\substack{c1,c2 \\ c1\neq c2}} \sum_{k1,k2} |\mathbf{y}^-_{c2,k2}|$ and we abuse notation by denoting $|\mathbf{y}_{c1,k1}|$ as the vector whose elements are the absolute values of the elements in $\mathbf{y}_{c1,k1}$.

Consider the related problem

$$\min_{\mathbf{y}_{c1,k1}} \|\mathbf{y}_{c1,k1} - \mathbf{q}_{c1,k1}\|_2^2 + \lambda_0((\mathbf{y}^+_{c1,k1})^T \mathbf{g}^+ + (\mathbf{y}^-_{c1,k1})^T \mathbf{g}^-) + \lambda_1 \|\mathbf{y}_{c1,k1}\|_1, \tag{46}$$

by taking the first order derivative w.r.t. $\mathbf{y}_{c1,k1}$ we have that

$$(\mathbf{y}_{c1,k1} - \mathbf{q}_{c1,k1}) + \lambda_1 \text{sign}(\mathbf{y}_{c1,k1}) + \lambda_0(\text{sign}(\mathbf{y}^+_{c1,k1}) \odot \mathbf{g}^+ + \text{sign}(\mathbf{y}^-_{c1,k1}) \odot \mathbf{g}^-) = \mathbf{0}, \tag{47}$$

take sign magnitude decomposition of $\mathbf{y}_{c1,k1} = \text{sign}(\mathbf{y}_{c1,k1}) \odot |\mathbf{y}_{c1,k1}|$ then we have

$$\begin{aligned} \text{sign}(\mathbf{y}_{c1,k1}) \odot |\mathbf{y}_{c1,k1}| - \text{sign}(\mathbf{q}_{c1,k1}) \odot |\mathbf{q}_{c1,k1}| + \\ \lambda_1 \text{sign}(\mathbf{y}_{c1,k1}) + \lambda_0(\text{sign}(\mathbf{y}^+_{c1,k1}) \odot \mathbf{g}^+ + \text{sign}(\mathbf{y}^-_{c1,k1}) \odot \mathbf{g}^-) = \mathbf{0}. \end{aligned} \tag{48}$$

Let the sign of $\mathbf{y}_{c1,k1}$, *i.e.* $\text{sign}(\mathbf{y}_{c1,k1})$ be equal to the sign of $\text{sign}(\mathbf{q}_{c1,k1})$, and Hadamard multiply from the left side by $\text{sign}(\mathbf{q}_{c1,k1})$ then we have

$$\begin{aligned} & |\mathbf{y}_{c1,k1}| - |\mathbf{q}_{c1,k1}| + \lambda_1 \mathbf{1} + \lambda_0(\text{sign}(\mathbf{q}_{c1,k1}) \odot \text{sign}(\mathbf{q}^+_{c1,k1}) \odot \mathbf{g}^+ + \\ & \text{sign}(\mathbf{q}_{c1,k1}) \odot \text{sign}(\mathbf{q}^-_{c1,k1}) \odot \mathbf{g}^-) = \mathbf{0}, \end{aligned} \tag{49}$$

note that $\text{sign}(\mathbf{q}_{c1,k1}) \odot \text{sign}(\mathbf{q}^+_{c1,k1}) = \text{sign}(\mathbf{q}^+_{c1,k1})$ and that $\text{sign}(\mathbf{q}_{c1,k1}) \odot \text{sign}(\mathbf{q}^-_{c1,k1}) = \text{sign}(-\mathbf{q}^-_{c1,k1})$, therefore we have

$$|\mathbf{y}_{c1,k1}| = |\mathbf{q}_{c1,k1}| - \lambda_1 \mathbf{1} - \lambda_0(\text{sign}(\mathbf{q}^+_{c1,k1}) \odot \mathbf{g}^+ + \text{sign}(-\mathbf{q}^-_{c1,k1}) \odot \mathbf{g}^-). \tag{50}$$

Define $\mathbf{G}^{c1,k1}$ as diagonal matrix with positive diagonal elements $G^{c1,k1}(m,m) = z(m), \forall m \in \mathcal{M}$ where:

$$\mathbf{z} = \text{sign}(\max(\mathbf{q}_{c1,k1}, \mathbf{0})) \odot \mathbf{g}^+ + \text{sign}(\max(-\mathbf{q}_{c1,k1}, \mathbf{0})) \odot \mathbf{g}^-, \tag{51}$$

and note that $\mathbf{G}^{c1,k1}\mathbf{1} = \mathbf{z}$ then

$$|\mathbf{y}_{c1,k1}| = |\mathbf{q}_{c1,k1}| - \lambda_1 \mathbf{1} - \lambda_0(\mathbf{G}^{c1,k1}\mathbf{1}), \tag{52}$$

since the magnitude might be only positive $|\mathbf{y}_{c1,k1}| = \max(|\mathbf{q}_{c1,k1}| - \lambda_1 \mathbf{1} - \lambda_0(\mathbf{G}^{c1,k1}\mathbf{1}), \mathbf{0})$. Therefore, the closed form solution to (46) is:

$$\mathbf{y}_{c1,k1} = \text{sign}(\mathbf{y}_{c1,k1}) \odot \max(|\mathbf{q}_{c1,k1}| - \lambda_1 \mathbf{1} - \lambda_0 \mathbf{G}^{c1,k1}\mathbf{1}, \mathbf{0}). \tag{53}$$

which completes the proof □

APPENDIX C.

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

$$\begin{aligned}\mathbf{y} =& \mathcal{T}(\mathbf{A}\mathbf{x}) = \max(\mathbf{A}\mathbf{x} - \boldsymbol{\tau}, \mathbf{0}) - \max(-\mathbf{A}\mathbf{x} - \boldsymbol{\tau}, \mathbf{0}),\\ \mathbf{q} =& \mathcal{T}(\mathbf{A}\mathbf{x}) = \max(\mathbf{A}\mathbf{g} - \boldsymbol{\tau}, \mathbf{0}) - \max(-\mathbf{A}\mathbf{g} - \boldsymbol{\tau}, \mathbf{0}),\end{aligned} \tag{54}$$

since

$$\operatorname{sign}(a)\max(|a| - b, 0) = \max(a - b, 0) - \max(-a - b, 0), \tag{55}$$

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

$$\begin{aligned}&\frac{\partial D_{\ell_1}^{\mathcal{P}^t}(\mathbf{x};\mathbf{g})}{\partial \mathbf{A}} =\\ &\frac{\partial(\max(\mathbf{A}\mathbf{x} - \boldsymbol{\tau}, \mathbf{0})^T \max(\mathbf{A}\mathbf{g} - \boldsymbol{\tau}, \mathbf{0}))}{\partial \mathbf{A}} + \frac{\partial(\max(-\mathbf{A}\mathbf{x} - \boldsymbol{\tau}, \mathbf{0})^T \max(-\mathbf{A}\mathbf{g} - \boldsymbol{\tau}, \mathbf{0}))}{\partial \mathbf{A}}\end{aligned} \tag{56}$$

we assume that the threshold parameter $\boldsymbol{\tau}$ is chosen such that the vector $|\mathbf{A}\mathbf{x}| - \boldsymbol{\tau}$ (or for any other $\mathbf{q}$, the vector $|\mathbf{A}\mathbf{q}| - \boldsymbol{\tau}$) has least one non-zero element, then

$$\frac{\partial(\max(\mathbf{A}\mathbf{x} - \boldsymbol{\tau}, \mathbf{0})^T \max(\mathbf{A}\mathbf{g} - \boldsymbol{\tau}, \mathbf{0}))}{\partial \mathbf{A}} = \max(\mathbf{A}\mathbf{g} - \boldsymbol{\tau}, \mathbf{0})\mathbf{x}^T + \max(\mathbf{A}\mathbf{x} - \boldsymbol{\tau}, \mathbf{0})\mathbf{g}^T, \tag{57}$$

and

$$\begin{aligned}&\frac{\partial(\max(-\mathbf{A}\mathbf{x} - \boldsymbol{\tau}, \mathbf{0})^T \max(-\mathbf{A}\mathbf{g} - \boldsymbol{\tau}, \mathbf{0}))}{\partial \mathbf{A}} =\\ &- \max(-\mathbf{A}\mathbf{g} - \boldsymbol{\tau}, \mathbf{0})\mathbf{x}^T - \max(-\mathbf{A}\mathbf{x} - \boldsymbol{\tau}, \mathbf{0})\mathbf{g}^T,\end{aligned} \tag{58}$$

combining (57) and (58) we have that

$$\begin{aligned}&\frac{\partial D_{\ell_1}^{\mathcal{P}^t}(\mathbf{x};\mathbf{g})}{\partial \mathbf{A}} =\\ &\max(\mathbf{A}\mathbf{g} - \boldsymbol{\tau}, \mathbf{0})\mathbf{x}^T + \max(\mathbf{A}\mathbf{x} - \boldsymbol{\tau}, \mathbf{0})\mathbf{g}^T -\\ &(\max(-\mathbf{A}\mathbf{g} + \boldsymbol{\tau}, \mathbf{0})\mathbf{x}^T + \max(-\mathbf{A}\mathbf{x} - \boldsymbol{\tau}, \mathbf{0})\mathbf{g}^T) =\\ &\mathbf{q}\mathbf{x}^T + \mathbf{y}\mathbf{g}^T\end{aligned} \tag{59}$$

where

$$\begin{aligned}y(m) &= \operatorname{sign}(\mathbf{a}_m^T\mathbf{x})\max(|\mathbf{a}_m^T\mathbf{x}| - \tau(m), 0)\\ q(m) &= \operatorname{sign}(\mathbf{a}_m^T\mathbf{g})\max(|\mathbf{a}_m^T\mathbf{g}| - \tau(m), 0),\end{aligned} \tag{60}$$

$\forall m \in \mathcal{M}$.

Similarity, note that for the model $\mathcal{P}_0^o = \{\mathbf{A}_o, \boldsymbol{\tau} = \mathbf{0}\}$ we have that

$$\begin{aligned}\mathbf{x}_o =& \mathcal{T}^{\mathcal{P}_0^o}(\mathbf{A}_o\mathbf{x}) =\\ &\operatorname{sign}(\mathbf{A}_o\mathbf{x}) \odot \max(|\mathbf{A}_o\mathbf{x}| - \mathbf{0}, \mathbf{0}) =\\ &\max(\mathbf{A}_o\mathbf{x} - \mathbf{0}, \mathbf{0}) - \max(-\mathbf{A}_o\mathbf{x} - \mathbf{0}, \mathbf{0}),\\ \mathbf{g}_o =& \mathcal{T}^{\mathcal{P}_0^o}(\mathbf{A}_o\mathbf{g}) =\\ &\operatorname{sign}(\mathbf{A}_o\mathbf{g}) \odot \max(|\mathbf{A}_o\mathbf{g}| - \mathbf{g}, \mathbf{0}) =\\ &\max(\mathbf{A}_o\mathbf{g} - \mathbf{0}, \mathbf{0}) - \max(-\mathbf{A}_o\mathbf{g} - \mathbf{0}, \mathbf{0}).\end{aligned} \tag{61}$$

The first order derivative of the divergence $D_{\ell_1}^{\mathcal{P}_0^o}(\mathbf{x};\mathbf{g})$ w.r.t $\mathbf{A}_o$ is:

$$\frac{\partial D_{\ell_1}^{\mathcal{P}_0^o}(\mathbf{x};\mathbf{g})}{\partial \mathbf{A}_o} = \mathbf{x}_o\mathbf{g}^T + \mathbf{g}_o\mathbf{x}^T, \tag{62}$$

note that at $\mathbf{A}_o = \mathbf{I}$, $\mathcal{P}_0^t = \mathcal{B}^N$ and we have that

$$\frac{\partial D_{\ell_1}^{\mathcal{B}^N}(\mathbf{x};\mathbf{g})}{\partial \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}} = \mathbf{x}\mathbf{g}^T + \mathbf{g}\mathbf{x}^T \tag{63}$$

Also for the model $\mathcal{P}_0^t = \{\mathbf{A}_t, \boldsymbol{\tau} = \mathbf{0}\}$ we have that

$$\begin{aligned}\mathbf{y}_t =& \mathcal{T}^{\mathcal{P}_0^t}(\mathbf{A}_t\mathbf{y}) = \\ & \operatorname{sign}(\mathbf{A}_t\mathbf{y}) \odot \max(|\mathbf{A}_t\mathbf{y}| - \mathbf{0}, \mathbf{0}) = \\ & \max(\mathbf{A}_t\mathbf{y} - \mathbf{0}, \mathbf{0}) - \max(-\mathbf{A}_t\mathbf{y} - \mathbf{0}, \mathbf{0}), \\ \mathbf{q}_t =& \mathcal{T}^{\mathcal{P}_0^t}(\mathbf{A}_t\mathbf{q}) = \\ & \operatorname{sign}(\mathbf{A}_t\mathbf{q}) \odot \max(|\mathbf{A}_t\mathbf{q}| - \mathbf{q}, \mathbf{0}) = \\ & \max(\mathbf{A}_t\mathbf{q} - \mathbf{0}, \mathbf{0}) - \max(-\mathbf{A}_t\mathbf{q} - \mathbf{0}, \mathbf{0}).\end{aligned} \tag{64}$$

The first order derivative of the divergence $D_{\ell_1}^{\mathcal{P}_0^t}(\mathbf{y};\mathbf{q})$ w.r.t $\mathbf{A}_t$ is:

$$\frac{\partial D_{\ell_1}^{\mathcal{P}_0^t}(\mathbf{y};\mathbf{q})}{\partial \mathbf{A}_t} = \mathbf{y}_t\mathbf{q}^T + \mathbf{q}_t\mathbf{y}^T, \tag{65}$$

note that at $\mathbf{A}_t = \mathbf{I}$, $\mathcal{P}_0^t = \mathcal{B}^M$ and we have that

$$\frac{\partial D_{\ell_1}^{\mathcal{B}^M}(\mathbf{y};\mathbf{q})}{\partial \mathbf{A}_t}|_{\mathbf{A}_o=\mathbf{I}} = \mathbf{y}\mathbf{q}^T + \mathbf{q}\mathbf{y}^T \tag{66}$$

Consider the following

$$\begin{aligned}\mathbf{A}\mathbf{g} = \mathbf{q} + \mathbf{z}_1/\mathbf{x}^T \\ \mathbf{A}\mathbf{x} = \mathbf{y} + \mathbf{z}_2/\mathbf{g}^T\end{aligned} \rightarrow \left\{\begin{aligned}\mathbf{A}\mathbf{x}\mathbf{g}^T = \mathbf{y}\mathbf{g}^T + \mathbf{z}_1\mathbf{g}^T \\ \mathbf{A}\mathbf{g}\mathbf{x}^T = \mathbf{q}\mathbf{x}^T + \mathbf{z}_2\mathbf{x}^T\end{aligned}\right. \\ \mathbf{A}(\mathbf{x}\mathbf{g}^T + \mathbf{g}\mathbf{x}^T) = \mathbf{y}\mathbf{g}^T + \mathbf{q}\mathbf{x}^T + \mathbf{z}_1\mathbf{g}^T + \mathbf{z}_2\mathbf{x}^T \tag{67}$$

where

$$\begin{aligned}\mathbf{z}_1 &= \mathbf{A}\mathbf{x} - \operatorname{sign}(\mathbf{A}\mathbf{x})\max(|\mathbf{A}\mathbf{x}| - \tau\mathbf{1}, \mathbf{0}) \\ \mathbf{z}_2 &= \mathbf{A}\mathbf{g} - \operatorname{sign}(\mathbf{A}\mathbf{g})\max(|\mathbf{A}\mathbf{g}| - \tau\mathbf{1}, \mathbf{0})\end{aligned} \tag{68}$$

A closer look at (67) reveals us that

$$\mathbf{A}\frac{\partial D_{\ell_1}^{\mathcal{B}^N}(\mathbf{x};\mathbf{g})}{\partial \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}} = \frac{\partial D_{\ell_1}^{\mathcal{P}}(\mathbf{x};\mathbf{g})}{\partial \mathbf{A}} + \boldsymbol{\delta}_o^{z_1,z_2} \tag{69}$$

where

$$\boldsymbol{\delta}_o^{z_1,z_2} = \mathbf{z}_1\mathbf{g}^T + \mathbf{z}_2\mathbf{x}^T \tag{70}$$

By similar construction applied to the rest of the pairs of data samples we have the result:

$$\mathbf{A}\left(\frac{\partial D_{\ell_1,c}^{\mathcal{B}^N}(\mathbf{X})}{\partial \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}} + \frac{\partial D_{\ell_1}^{\mathcal{B}^N}(\mathbf{X})}{\partial \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}\right) = \frac{\partial D_{\ell_1,c}^{\mathcal{P}}(\mathbf{X})}{\partial \mathbf{A}} + \frac{\partial D_{\ell_1}^{\mathcal{P}}(\mathbf{X})}{\partial \mathbf{A}} + \boldsymbol{\delta}_o\square \tag{71}$$

Note that

$$\begin{aligned}\mathbf{A}\mathbf{x} = \mathbf{y} + \mathbf{z}_1/\mathbf{q}^T \\ \mathbf{A}\mathbf{g} = \mathbf{q} + \mathbf{z}_2/\mathbf{y}^T\end{aligned} \rightarrow \left\{\begin{aligned}\mathbf{A}\mathbf{x}\mathbf{q}^T = \mathbf{y}\mathbf{q}^T + \mathbf{z}_1\mathbf{q}^T \\ \mathbf{A}\mathbf{g}\mathbf{y}^T = \mathbf{q}\mathbf{y}^T + \mathbf{z}_2\mathbf{y}^T\end{aligned}\right. \\ \mathbf{A}(\mathbf{x}\mathbf{y}^T + \mathbf{q}\mathbf{g}^T) = \mathbf{y}\mathbf{q}^T + \mathbf{q}\mathbf{y}^T + \mathbf{z}_1\mathbf{q}^T + \mathbf{z}_2\mathbf{y}^T, \tag{72}$$

where

$$\begin{aligned}\mathbf{z}_1 &= \mathbf{A}\mathbf{x} - \operatorname{sign}(\mathbf{A}\mathbf{x})\max(|\mathbf{A}\mathbf{x}| - \tau\mathbf{1}, \mathbf{0}) \\ \mathbf{z}_2 &= \mathbf{A}\mathbf{g} - \operatorname{sign}(\mathbf{A}\mathbf{g})\max(|\mathbf{A}\mathbf{g}| - \tau\mathbf{1}, \mathbf{0})\end{aligned} \tag{73}$$

A closer look at (72) reveals us

$$\mathbf{A}\left(\frac{\partial D^{\mathcal{P}}_{\ell_1}(\mathbf{x};\mathbf{g})}{\partial \mathbf{A}}\right)^T = \frac{\partial D^{\mathcal{B}^M}_{\ell_1}(\mathbf{y};\mathbf{q})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} + \boldsymbol{\delta}_t^{z1,z2}. \tag{74}$$

where

$$\boldsymbol{\delta}_t^{z_1,z_2} = \mathbf{z}_1\mathbf{q}^T + \mathbf{z}_2\mathbf{y}^T \tag{75}$$

By similar construction applied to the rest of the pairs of data samples we have the result:

$$\mathbf{A}\left(\frac{\partial D^{\mathcal{P}^t}_{\ell_1,c}(\mathbf{X})}{\partial \mathbf{A}} + \frac{\partial D^{\mathcal{P}^t}_{\ell_1}(\mathbf{X})}{\partial \mathbf{A}}\right)^T = \left(\frac{\partial D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} + \frac{\partial D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}\right) + \boldsymbol{\delta}_t\square \tag{76}$$

Note that

$$\begin{array}{c} \mathbf{A}\mathbf{x} = \mathbf{y} + \mathbf{z}_1/\mathbf{z}_2^T \\ \mathbf{A}\mathbf{g} = \mathbf{q} + \mathbf{z}_2/\mathbf{z}_1^T \end{array} \rightarrow \left\{ \begin{array}{l} \mathbf{A}\mathbf{x}\mathbf{z}_2^T = \mathbf{y}\mathbf{z}_2^T + \mathbf{z}_1\mathbf{z}_2^T \\ \mathbf{A}\mathbf{g}\mathbf{z}_1^T = \mathbf{q}\mathbf{z}_1^T + \mathbf{z}_2\mathbf{z}_1^T \end{array} \right. \\ \mathbf{A}(\mathbf{x}\mathbf{z}_2^T + \mathbf{q}\mathbf{z}_1^T) = \mathbf{y}\mathbf{z}_2^T + \mathbf{q}\mathbf{z}_1^T + \mathbf{z}_1\mathbf{z}_2^T + \mathbf{z}_2\mathbf{z}_1^T, \tag{77}$$

where

$$\begin{array}{l} \mathbf{z}_1 = \mathbf{A}\mathbf{x} - \text{sign}(\mathbf{A}\mathbf{x})\max(|\mathbf{A}\mathbf{x}| - \tau\mathbf{1}, \mathbf{0}) \\ \mathbf{z}_2 = \mathbf{A}\mathbf{g} - \text{sign}(\mathbf{A}\mathbf{g})\max(|\mathbf{A}\mathbf{g}| - \tau\mathbf{1}, \mathbf{0}) \end{array} \tag{78}$$

A closer look at (77) reveals us

$$\mathbf{A}(\boldsymbol{\delta}_o^{z_1,z_2})^T = \frac{\partial D^{\mathcal{B}^M}_{\ell_1}(\mathbf{z}_1;\mathbf{z}_2)}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} + (\boldsymbol{\delta}_t^{z1,z2})^T. \tag{79}$$

where

$$\boldsymbol{\delta}_t^{z_1,z_2} = \mathbf{z}_1\mathbf{q}^T + \mathbf{z}_2\mathbf{y}^T \tag{80}$$

By similar construction applied to the rest of the pairs of data samples we have the result:

$$\mathbf{A}\boldsymbol{\delta}_o{}^T = \left(\frac{\partial D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Z})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}} + \frac{\partial D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Z})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}\right) + \boldsymbol{\delta}_t^T\square \tag{81}$$

APPENDIX D.

The result in (16) decomposes on the contributing components for similarity and the contributing components for dissimilarity, *i.e.*, $\frac{\partial \mathcal{J}_{\ell_1}(\mathbf{AX})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}|_s - \frac{\partial \mathcal{J}_{\ell_1}(\mathbf{AX})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}|_d = \frac{\partial \mathcal{J}_{\ell_1}(\mathbf{Y})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}|_s - \frac{\partial \mathcal{J}_{\ell_1}(\mathbf{Y})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}|_d + \boldsymbol{\xi}_{c,d} - \boldsymbol{\xi}_{c,d} + \boldsymbol{\xi}_s - \boldsymbol{\xi}_d$. Moreover, w.r.t. the similarity concentrations we have the following splitting $Tr\{\frac{\partial D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{AX})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}|_s\} = D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{Y}) + Tr\{\boldsymbol{\xi}_{c,s}\} = D^{\mathcal{P}}_{\ell_1,c}(\mathbf{X}) + Tr\{\boldsymbol{\xi}_{c,s}\}$ and $Tr\{\frac{\partial D^{\mathcal{B}^M}_{\ell_1}(\mathbf{AX})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}|_s\} = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{Y}) + Tr\{\boldsymbol{\xi}_s\} = D^{\mathcal{P}}_{\ell_1,c}(\mathbf{X}) + Tr\{\boldsymbol{\xi}_s\}$. Therefore, we have the following bounds

$$\begin{array}{l} a: Tr\{\mathbf{A}\frac{\partial D^{\mathcal{B}^N}_{\ell_1,c}(\mathbf{X})}{\partial \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}\mathbf{A}^T\} \leq D^{\mathcal{P}}_{\ell_1,c}(\mathbf{X}) \leq Tr\{\frac{\partial D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{AX})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}|_s\} = D^{\mathcal{B}^M}_{\ell_1,c}(\mathbf{AX}) \\ b: Tr\{\mathbf{A}\frac{\partial D^{\mathcal{B}^N}_{\ell_1}(\mathbf{X})}{\partial \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}\mathbf{A}^T\} \leq D^{\mathcal{P}}_{\ell_1}(\mathbf{X}) \leq Tr\{\frac{\partial D^{\mathcal{B}^M}_{\ell_1}(\mathbf{AX})}{\partial \mathbf{A}_t}|_{\mathbf{A}_t=\mathbf{I}}|_s\} = D^{\mathcal{B}^M}_{\ell_1}(\mathbf{AX}) \end{array} \tag{82}$$

Note that $c: \lambda_{min}(\mathbf{A}^T\mathbf{A})Tr\{\frac{\partial D^{\mathcal{B}^N}_{\ell_1,c}(\mathbf{X})}{\partial \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}\} \leq Tr\{\mathbf{A}\frac{\partial D^{\mathcal{B}^N}_{\ell_1,c}(\mathbf{X})}{\partial \mathbf{A}_o}|_{\mathbf{A}_o=\mathbf{I}}\mathbf{A}^T\}$ where $\lambda_{min}(\mathbf{A}^T\mathbf{A})$ is the minimum singular value to the matrix $\mathbf{A}^T\mathbf{A}$. Taking the logarithm of the ratio $\frac{D^{\mathcal{P}}_{\ell_1,c}(\mathbf{X})}{D^{\mathcal{P}}_{\ell_1}(\mathbf{X})+\epsilon}$ and using the bounds a, b and c we arrive at the desired result $\square$

REFERENCES

Shun-ichi Amari. Information geometry and its applications: Survey. In *GSI*, volume 8085 of *Lecture Notes in Computer Science*, pp. 3. Springer, 2013.

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012. URL http://arxiv.org/abs/1206.5538.

Sijia Cai, Wangmeng Zuo, Lei Zhang, Xiangchu Feng, and Ping Wang. Support vector guided dictionary learning. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, pp. 624–639, 2014.

Mehrdad J. Gangeh, Ahmed K. Farahat, Ali Ghodsi, and Mohamed S. Kamel. Supervised dictionary learning and sparse representation-a review. *CoRR*, abs/1502.05928, 2015. URL http://arxiv.org/abs/1502.05928.

A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:643–660, 2001.

Huimin Guo, Zhuolin Jiang, and Larry S. Davis. Discriminative dictionary learning with pairwise constraints. In *Computer Vision - ACCV 2012 - 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I*, pp. 328–342, 2012.

Huimin Guo, Zhuolin Jiang, and Larry S. Davis. *Discriminative Dictionary Learning with Pairwise Constraints*, pp. 328–342. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-37331-2. doi: 10.1007/978-3-642-37331-2_25. URL https://doi.org/10.1007/978-3-642-37331-2_25.

Rui Jiang, Hong Qiao, and Bo Zhang. Efficient fisher discrimination dictionary learning. *Signal Process.*, 128(C):28–39, November 2016. ISSN 0165-1684. doi: 10.1016/j.sigpro.2016.03.013. URL http://dx.doi.org/10.1016/j.sigpro.2016.03.013.

Zhuolin Jiang, Zhe Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pp. 1697–1704, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-1-4577-0394-2. doi: 10.1109/CVPR.2011.5995354. URL http://dx.doi.org/10.1109/CVPR.2011.5995354.

Zhuolin Jiang, Zhe Lin, and Larry S. Davis. Label consistent K-SVD: learning a discriminative dictionary for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2651–2664, 2013. doi: 10.1109/TPAMI.2013.88. URL http://dx.doi.org/10.1109/TPAMI.2013.88.

Dimche Kostadinov, Sviatoslav Volshinovsky, and Sohrab Ferdowsi. Learning non-structured, overcomplete and sparsifying transform. In *Lisbon, Portugal, SPARS 2017*, June 2017.

Kenneth Kreutz-Delgado, Joseph F. Murray, Bhaskar D. Rao, Kjersti Engan, Te-Won Lee, and Terrence J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Comput.*, 15(2):349–396, February 2003. ISSN 0899-7667. doi: 10.1162/089976603762552951. URL http://dx.doi.org/10.1162/089976603762552951.

Yann Lecun and Corinna Cortes. The MNIST database of handwritten digits. URL http://yann.lecun.com/exdb/mnist/.

Yann LeCun, Fu Jie Huang, and Léon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR'04, pp. 97–104, Washington, DC, USA, 2004. IEEE Computer Society.

Yann LeCun, David G. Lowe, Jitendra Malik, Jim Mutch, Pietro Perona, and Tomaso Poggio. Object Recognition, Computer Vision, and the Caltech 101: A Response to Pinto et al. Technical report, March 2008.

Yang Liu, Wei Chen, Qingchao Chen, and Ian J. Wassell. Support discrimination dictionary learning for image classification. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, pp. 375–390, 2016.

Julien Mairal, Francis R. Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Supervised dictionary learning. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pp. 1033–1040, 2008.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 689–696, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553463. URL `http://doi.acm.org/10.1145/1553374.1553463`.

A. Martínez and R. Benavente. The ar face database. Technical Report 24, Computer Vision Center, Jun 1998. URL "`http://www.cat.uab.cat/Public/Publications/1998/MaB1998`".

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (coil-20. Technical report, 1996.

D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pp. 2161–2168, June 2006. **oral presentation**.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, January 2014. ISSN 2167-3888.

I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3501–3508, June 2010. doi: 10.1109/CVPR.2010.5539964.

Saiprasad Ravishankar and Yoram Bresler. Learning sparsifying transforms for image processing. In *19th IEEE International Conference on Image Processing, ICIP 2012, Lake Buena Vista, Orlando, FL, USA, September 30 - October 3, 2012*, pp. 681–684. IEEE, 2012. ISBN 978-1-4673-2534-9. doi: 10.1109/ICIP.2012.6466951. URL `http://dx.doi.org/10.1109/ICIP.2012.6466951`.

Saiprasad Ravishankar and Yoram Bresler. Doubly sparse transform learning with convergence guarantees. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, pp. 5262–5266. IEEE, 2014. doi: 10.1109/ICASSP.2014.6854607. URL `http://dx.doi.org/10.1109/ICASSP.2014.6854607`.

Ron Rubinstein and Michael Elad. Dictionary learning for analysis-synthesis thresholding. *IEEE Trans. Signal Processing*, 62(22):5962–5972, 2014.

Ron Rubinstein, Alfred M. Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

Ron Rubinstein, Tomer Peleg, and Michael Elad. Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model. *IEEE Trans. Signal Processing*, 61(3):661–677, 2013.

Sumit Shekhar, Vishal M. Patel, and Rama Chellappa. Analysis sparse coding models for image-based classification. In *2014 IEEE International Conference on Image Processing, ICIP 2014, Paris, France, October 27-30, 2014*, pp. 5207–5211. IEEE, 2014. ISBN 978-1-4799-5751-4. doi: 10.1109/ICIP.2014.7026054. URL `http://dx.doi.org/10.1109/ICIP.2014.7026054`.

Ali Taalimi, Shahab Ensafi, Hairong Qi, Shijian Lu, Ashraf A. Kassim, and Chew Lim Tan. Multimodal dictionary learning and joint sparse representation for hep-2 cell classification. pp. 308–315, 2015.

T. H. Vu and V. Monga. Learning a low-rank shared dictionary for object classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 4428–4432, Sept 2016a. doi: 10.1109/ICIP.2016.7533197.

T. H. Vu, H. S. Mousavi, V. Monga, U. K. A. Rao, and G. Rao. Dfdl: Discriminative feature-oriented dictionary learning for histopathological image classification. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 990–994, April 2015. doi: 10.1109/ISBI.2015.7164037.

Tiep Huu Vu and Vishal Monga. Fast low-rank shared dictionary learning for image classification. *CoRR*, abs/1610.08606, 2016b. URL `http://arxiv.org/abs/1610.08606`.

Tiep Huu Vu and Vishal Monga. Fast low-rank shared dictionary learning for image classification. *CoRR*, abs/1610.08606, 2016c. URL `http://arxiv.org/abs/1610.08606`.

Yong Xu, Yuping Sun, Yuhui Quan, and Bo Zheng. Discriminative structured dictionary learning with hierarchical group sparsity. *Comput. Vis. Image Underst.*, 136(C):59–68, July 2015. ISSN 1077-3142. doi: 10.1016/j.cviu.2015.01.006. URL `http://dx.doi.org/10.1016/j.cviu.2015.01.006`.

M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *2011 International Conference on Computer Vision*, pp. 543–550, Nov 2011a. doi: 10.1109/ICCV.2011.6126286.

Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang. Fisher discrimination dictionary learning for sparse representation. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pp. 543–550, Washington, DC, USA, 2011b. IEEE Computer Society. ISBN 978-1-4577-1101-5. doi: 10.1109/ICCV.2011.6126286. URL `http://dx.doi.org/10.1109/ICCV.2011.6126286`.