Learning Discrimination Specific, Self-Collaborative and Nonlinear Model

Dimche Kostadinov, Behrooz Razeghi, Sviatoslav Voloshynovskiy and Sohrab Ferdowsi
Stochastic Information Processing Group, web: http://sip.unige.ch
Department of Computer Science, *University of Geneva, Switzerland*
rute de Drize 7, 1227 Carouge, Geneva, Switzerland
{Dimche.Kostadinov, Behrooz.Razeghi, Svolos, Sohrab.Ferdowsi}@unige.ch

***Abstract*—This paper presents a novel nonlinear transform model for learning of collaboration structured, discriminative and sparse representations. The idea is to model a collaboration corrective functionality between multiple nonlinear transforms in order to reduce the uncertainty in the estimate. The focus is on the joint estimation of a data-adaptive nonlinear transforms (NTs) that take into account a collaboration component w.r.t. a discrimination target. The joint model includes minimum information loss, collaboration corrective and discriminative priors. The model parameters are learned by minimizing the empirical negative log likelihood of the model, where we propose an efficient solution by an iterative, coordinate descend algorithm. Numerical experiments validate the potential of the learning principle. The preliminary results show advantages in comparison to the state-of-the-art methods, w.r.t. the learning time, the discriminative quality and the recognition accuracy.**

***Index Terms*—unsupervised feature learning, target specific self-collaborative model, discriminative min-max prior, joint learning of nonlinear transforms**

I. Introduction

In the recent years, the area of machine learning has had significant progress and advances. Various algorithms in different applications showed excellent results. Crucial to many approaches is the estimation/learning of task-relevant, useful and information preserving representation.

To differentiate the data that originates from different groups, many unsupervised learning methods [24], [35], [1], [18], [36], [3], [8], [11] were proposed. Their primary target is to describe and identify the underlining explanatory groups within the data with (or without) data priors. Usually, a data representation expressed with respect to the groups is used as an unsupervised feature.

Many discriminative descriptions were offered by the sparse (or structured sparse) models [4], [15], [14], [2],[16], the discriminative clustering approaches [42], [18] and the discrimination dictionary learning methods [28], [43], [38] and [37], where it is assumed that the true data exhibits a form of sparse structure.

However, so far, to the best of our knowledge, a joint discrimination centered, collaboration structured and sparse modeling was not explored. Due to the ambiguities in specifying a notion of discrimination and task focused collaboration, the related learning problem is challenging. The main open issues are the data model and the appropriate priors, which delimit the problem formulation and the definition of a suitable objective.

This research has been partially supported by the ERA-Net project ID_IoT No 20CH21_167534.

A. Joint Modeling of Nonlinear Transforms with Target Specific Self-Collaboration

In general, for any modeling usually we assume an error distribution that (1) reflects to the model correctness w.r.t. the true data distribution and the task at hand and (2) is significant for the robustness in the estimate. At the same time the right or wrong error assumption is crucial since it leads to accumulation or removal of uncertainties related to the target-specific goal.

– *Motivations:* Under a priori unknown error distribution w.r.t. certain task, a single estimate might have high variability. Therefore, a joint model of multiple transforms, where a relation between the errors of the transform representations is addressed might be more suitable. Since even if the errors in the transform representation alone have high variability, the joint composition will have the possibility to compensate for this variability.

– *Discriminative Self-Collaboration Model:* In this paper, we propose a novel model for joint learning of multiple nonlinear transforms parametrized by linear maps and element-wise nonlinearities. Instead of focusing on a particular error distribution per transform we focus on explicitly modeling of a relationship between the transform errors. The modeling is towards a target-specific goal expressed through the self-collaboration discriminative and minimum information loss priors. In this scene, we introduce a self-collaboration functionality, which to the best of our knowledge is first of this kind that extends and generalizes the sparsifying transform model [6], [31] and [29]. The proposed model offers several advantages, including (i) specification of one or combination of arbitrary goals, (ii) structuring per target-specific goals, (iii) supervised, unsupervised and semi-supervised centric collaboration and (iv) parallel and distributed parameter learning.

– *Learning Strategy:* Given observed data, the model parameters are learned by minimizing the empirical approximation to the expected negative log likelihood of the model. The learning target is the expected negative log likelihood of the collaboration component and the discriminative prior that gives interpretation about the models empirical risk and unfolds it as optimization cost that has to be minimized during learning.

II. Sparse Modeling and Related Work

In the subsequent subsections, we first describe the common sparse models, including the sparsifying transform model [32], [31] and [30] that is basis to the model that we propose, then we give the related work in the line of discriminative and sparse representations.

A. 2.1. Sparse Models

In the following, we introduce the primary sparse models.

– *Synthesis Model:* A *synthesis model* [6] and [31] (or regression model with sparsity regularized penalty) synthesizes a data sample $\mathbf{x}_{c,k} \in \Re^N$ as an approximation by a sparse linear combination $\mathbf{y}_{c,k} \in \Re^M$ ($\|\mathbf{y}_{c,k}\|_0 << M$), of a few vectors $\mathbf{d}_m \in \Re^N$, from a dictionary $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_M] \in \Re^{N \times M}$, *i.e.*, $\mathbf{x}_{c,k} = \mathbf{D}\mathbf{y}_{c,k} + \mathbf{z}_{c,k}$, where $\mathbf{z}_{c,k} \in \Re^N$ denotes the error vector defined in the original data domain.

– *Analysis Model:* It uses a dictionary $\mathbf{\Omega} \in \Re^{M \times N}$ with $M > N$ to *analyze* the data $\mathbf{x}_{c,k} \in \Re^N$. This model assumes that the product of $\mathbf{\Omega}$ and $\mathbf{x}_{c,k}$ is sparse, *i.e.*, $\mathbf{y}_{c,k} = \mathbf{\Omega}\mathbf{x}_{c,k}$ with $\|\mathbf{y}_{c,k}\|_0 = M - s$, where $0 \leq s \leq M$ is the number of zeros in $\mathbf{y} \in \Re^M$ [33] and [10]. The vector $\mathbf{y}_{c,k}$ is the analysis sparse representation of the data $\mathbf{x}_{c,k}$ w.r.t. $\mathbf{\Omega}$.

– *Transform Model:* In contrast to the synthesis model and similar to the analysis model [6] [29], [30] and [17], the *sparsifying transform model* does not target explicitly the data reconstruction. 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 $\mathbf{y}$ is sparse $\|\mathbf{y}_{c,k}\|_0 << M$, and the error vector $\mathbf{z}_{c,k}$ is defined in the transform domain.

B. 2.2. Structured Discrimination Constraints

To address the estimation of discriminative sparse representations (in a supervised or unsupervised setup) the previous models usually are learned with a constraints. The commonly used penalties are set on (i) the dictionary, (ii) the sparse representation (*e.g.*, pairwise similarity, encoding w.r.t. a graph and structured sparsity) and (iii) the cost w.r.t. a classifier. In the following, we indicate the primary portion of related work.

– *Structured Sparsity Methods:* The structured sparsity models were widely used in many practical problems, including model-based compressive sensing [4], signal processing [6], [29], [30] and [10], computer vision [15], bio-informatics [39] and recommendation systems [16]. In this paper, we use structuring w.r.t. a targeted collaboration to reduce the uncertainty in the estimate w.r.t. the discriminative properties of the NT representations.

– *Discriminative Dictionary Learning* (*DDL*): Discrimination constraints were mainly defined by exploiting labels. The class of algorithms is known as discriminative dictionary learning methods (DDL) [28], [43], [38] and [37]. Our approach addresses the unsupervised case with discrimination constraints.

– *Discriminative Clustering:* In [42], clustering with maximum margin constraints was proposed. The authors in [1] proposed linear clustering based on a linear discriminative cost function with convex relaxation. In [18] regularized information maximization was proposed and simultaneous clustering and classifier training was preformed. The above methods rely on kernels and have high computational complexity. The proposed method learns to reduce or extend dimensionality through a transform that enforces discrimination.

– *Auto-encoders:* The single layer auto-encoder [3] and its denoising extension [36] consider robustness to noise and reconstruction. While the idea is to encode and decode the data using a reconstruction loss, an explicit constraint that enforces discrimination is not addressed.

III. Paper Organization and Notations

A. Paper Organization

Section 4 introduces and explains the joint model and the priors. Sections 5 reveals the learning target, identifies the empirical risk and presents the problem formulation. Section 6 proposes a solution using an iterative, alternating algorithm. Section 7 provides the numerical experiments and evaluation, and Section 8 concludes the paper.

B. Notations

A scalar, vector and matrix are denoted using standard, lower bold and upper bold case symbols as x, $\mathbf{x}$ and $\mathbf{X}$, respectively. A set of L sets of data representations is denoted as $\mathbf{Y} = [\mathbf{Y}_1, ..., \mathbf{Y}_L]$, where the set of data representations $\mathbf{Y}_l = [\mathbf{y}_{l,\{1,1\}}, ..., \mathbf{y}_{l,\{C,K\}}] \in \Re^{M \times CK}$. For every $l \in \{1, ..., L\}$, every class $c \in \mathcal{C} = \{1, ..., C\}$ has K samples, *i.e.*, $[\mathbf{y}_{l,\{c,1\}}, ..., \mathbf{y}_{l,\{c,K\}}] \in \Re^{M \times K}$. The set of L components for index $\{c, k\}$ is denoted as $\mathbf{Y}_{\{c,k\}} = [\mathbf{y}_{1,\{c,k\}}, ..., \mathbf{y}_{L,\{c,k\}}]$. The ℓ_p-norm, the Hadamard product and element-wise division are denoted as $\|.\|_p$, $\odot$ and $\oslash$, respectively.

IV. Target Specific Uncertainty Reduction Model

Our model cores around three elements: (i) a *self-collaboration nonlinear modeling*, (ii) *an unsupervised and collaborative discriminative prior* described using a (iii) *min-max cost* which includes a formally defined notion for similarity and dissimilarity contributions.

The model describes a generalized structured nonlinearity of L data representations $\mathbf{y}_{l,\{c,k\}}, \forall l \in \{1, ..., L\}$ that explain the data $\mathbf{x}_{c,k}$ with collaborative uncertainty reduction term and priors. The joint model expresses as $p(\mathbf{x}_{c,k}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}, \mathbf{A}) = p(\mathbf{x}_{c,k}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}|\mathbf{A})p(\mathbf{A})$, where we assume that $p(\mathbf{x}_{c,k}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}|\mathbf{A}) = p(\mathbf{x}_{c,k}|\mathbf{Y}_{\{c,k\}}, \mathbf{A})p(\boldsymbol{\theta}, \mathbf{Y}_{\{c,k\}})$.

A. Joint Modeling with Collaboration

We model multiple nonlinear transforms with a collaboration component as follows:

$$p(\mathbf{x}_{c,k}|\mathbf{Y}_{\{c,k\}}, \mathbf{A}) \propto \prod_l \exp(-\frac{1}{\beta_0}(\mathbf{z}_{l,\{c,k\}}^T \mathbf{z}_{l,\{c,k\}} + f_{TSC}(\mathbf{z}_{l,\{c,k\}}, g_A(\mathbf{Z}_{\{c,k\}\backslash l})))), \quad (1)$$

where $\mathbf{Y}_{\{c,k\}} = [\mathbf{y}_{1,\{c,k\}}, ..., \mathbf{y}_{L,\{c,k\}}]$, $\mathbf{A} = [\mathbf{A}_1, ..., \mathbf{A}_L]$, $\mathbf{Z}_{\{c,k\}\backslash l} = [\mathbf{z}_{1,\{c,k\}}, ..., \mathbf{z}_{l-1,\{c,k\}}, \mathbf{z}_{l+1,\{c,k\}}, ..., \mathbf{z}_{L,\{c,k\}}]$. The term $f_{TSC}(\mathbf{z}_{l,\{c,k\}}, g_A(\mathbf{Z}_{\{c,k\}\backslash l})) : \Re^M \times \Re^M \rightarrow \Re$ denotes a *target specific collaboration* function, that we define it as $\mathbf{z}_{l,\{c,k\}}^T g_A(\mathbf{Z}_{\{c,k\}\backslash l})$ and $g_A(\mathbf{Z}_{\{c,k\}\backslash l}) : \Re^M \times ... \times \Re^M \rightarrow \Re^M$

denotes the collaboration aggregation function, that we define as $g_A(\mathbf{Z}_{\{c,k\}\setminus l}) = \sum_{l1\neq l} \mathbf{z}_{l1,\{c,k\}} \in \Re^M$. The model (1) assumes that the data sample $\mathbf{x}_{c,k}$ indexed by k from group c is approximately sparsifiable under any of the linear transforms $\mathbf{A}_l \in \Re^{M\times N}$, $l \in \{1,...,L\}$, the target specific collaboration function f_{TSC} and the collaboration aggregation function g_A, *i.e.*, $\mathbf{A}_l\mathbf{x}_{c,k} = \mathbf{y}_{l,\{c,k\}} + \mathbf{v}_{l,\{c,k\}} + \sum_{l1\neq l} \mathbf{z}_{l1,\{c,k\}}$, where $\mathbf{y}_{l,\{c,k\}} \in \Re^M$ is the sparse representation, $\mathbf{v}_{l,\{c,k\}} \in \Re^M$ is the corrected sparsifying error vector and $\mathbf{z}_{l,\{c,k\}} = \mathbf{A}_l\mathbf{x}_{c,k} - \mathbf{y}_{l,\{c,k\}}$.

– *Self-Collaboration:* The uncertainty reduction by (1) is w.r.t. a target described by the relation on the error terms $\mathbf{z}_{l1,\{c,k\}}$. The terms $\mathbf{z}^T_{l,\{c,k\}} \sum_{l1\neq l} \mathbf{z}_{l1,\{c,k\}}$ act as a self-regularization that tries to compensate for the unknown distribution of $\mathbf{z}_{l,\{c,k\}}$ by reducing the uncertainty w.r.t. the affine combination $\mathbf{c}_{l,\{c,k\}} = \sum_{l1\neq l} \mathbf{z}_{l1,\{c,k\}}$ of the rest of unknown distributions for $\mathbf{z}_{l1,\{c,k\}}$. The goal is to allow the individual $\mathbf{z}_{l,\{c,k\}}$ to describe only a portion of the targeted, but, unknown probability distribution of $\mathbf{z}_{l,\{c,k\}}$ in order to reduce the uncertainty and give a reliable and robust estimate w.r.t. a composition and aggregation function that in our case is simply a concatenation $[\mathbf{y}_{1,\{c,k\}},...,\mathbf{y}_{L,\{c,k\}}]$.

– *Prior on* $\mathbf{z}_l$: The transform representation $\mathbf{y}_{l,\{c,k\}} = \mathcal{T}_{\mathcal{P}_{l,c}}(\mathbf{x}_{c,k})$ takes into account a nonlinearity and $\mathbf{A}_l\mathbf{x}_{c,k} - \mathbf{c}_{l,c}$ is only seen as its linear approximation. In the simplest form we model $\mathbf{z}_{l,\{c,k\}}$ by $p(\mathbf{x}_{c,k}|\mathbf{y}_{l,\{c,k\}}, \mathbf{A}_l)$ to be a Gausian distributed. Additional knowledge about $\mathbf{z}_{l,\{c,k\}}$ can be used to model $p(\mathbf{x}_{c,k}|\mathbf{y}_{l,\{c,k\}}, \mathbf{A}_l)$.

– *Prior on* $\mathbf{A}_l$: Additionally, we have a prior on $\mathbf{A}_l$ that penalizes the information loss in order to avoid trivially unwanted matrices $\mathbf{A}_l$, *i.e.*, matrices that have repeated or zero rows. This prior measure is denoted as $\Omega(\mathbf{A}_l) = (\frac{1}{\beta_{l,3}}\|\mathbf{A}_l\|_F^2 + \frac{1}{\beta_{l,4}}\|\mathbf{A}_l\mathbf{A}_l^T - \mathbf{I}\|_F^2 - \frac{1}{\beta_{l,5}}\log|\det \mathbf{A}_l^T\mathbf{A}_l|)$. The terms $\frac{1}{\beta_{l,3}}\|\mathbf{A}_l\|_F^2 + \frac{1}{\beta_{l,4}}\|\mathbf{A}_l\mathbf{A}_l^T - \mathbf{I}\|_F^2 - \frac{1}{\beta_{l,5}}\log|\det \mathbf{A}_l^T\mathbf{A}_l|$ are used to regularize the conditioning and the expected coherence of $\mathbf{A}_l$ (for more details please see [17]).

B. Self-Collaboration Discriminative Prior

A joint probability expresses the *unsupervised discriminative prior* as:

$$p(\boldsymbol{\theta}, \mathbf{Y}_{\{c,k\}}) = \prod_l p(\boldsymbol{\theta}|\mathbf{y}_{l,\{c,k\}})p(\mathbf{y}_{l,\{c,k\}}), \tag{2}$$

and it allows explicit modeling of (i) dependences between $\boldsymbol{\theta}$ and $\mathbf{y}_{l,\{c,k\}}$ or modeling of (ii) dependences between $\boldsymbol{\theta}_l$ and $\mathbf{y}_{l,\{c,k\}}$, where only the relation between $\boldsymbol{\theta}_l$ and $\mathbf{y}_{l,\{c,k\}}$ is considered.

In this paper we address independent modeling per $\boldsymbol{\theta}_l$ and $\mathbf{y}_{l,\{c,k\}}$ and consider that (2) has the form as:

$$p(\boldsymbol{\theta}, \mathbf{Y}_{\{c,k\}}) = \prod_l p(\boldsymbol{\theta}_l|\mathbf{y}_{l,\{c,k\}})p(\mathbf{y}_{l,\{c,k\}}), \tag{3}$$

and that $p(\boldsymbol{\theta}_l|\mathbf{y}_{l,\{c,k\}}) \propto \exp(-l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}}))$, where $l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}})$ is a discriminative measure for similarity contributions over the parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_L\}$ and $p(\mathbf{y}_{l,\{c,k\}}) \propto \exp(-\frac{\|\mathbf{y}_{l,\{c,k\}}\|_1}{\beta_{l,1}})$ is a sparisty inducing prior.

It is assumed that $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_L\}$ where $\boldsymbol{\theta}_l = \{\boldsymbol{\theta}_{l,1}, \boldsymbol{\theta}_{l,2}\} = \{\{\boldsymbol{\tau}_{l,1},...,\boldsymbol{\tau}_{l,C1}\}, \{\boldsymbol{\nu}_{l,1},...,\boldsymbol{\nu}_{l,C2}\}\}$ cover similarity and dissimilarity regions in the transform space and take into account a form of collaboration.

The use of the measure $l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}})$ allows us to introduce a concept that relies on collaboration corrective to covering of similarity and dissimilarity regions by taking into account a relations between $\boldsymbol{\theta}_l$ and $\mathbf{Y}_{\{c,k\}}$.

– *Prior Measures:* To define $l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}})$ it is assumed that (i) $p(\boldsymbol{\theta}_l) = p(\boldsymbol{\theta}_{l,1})p(\boldsymbol{\theta}_{l,2}) = \prod_{c1} p(\boldsymbol{\tau}_{l,c1}) \prod_{c2} p(\boldsymbol{\nu}_{l,c2})$, (ii) $l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}})$ is determined by relations on the *support intersection* between $\mathbf{y}_{l,\{c,k\}}$, $\boldsymbol{\tau}_{l,c1}$ and $\boldsymbol{\nu}_{l,c2}$, and (iii) the description is decomposable w.r.t. $\boldsymbol{\tau}_{l,c1}$ and $\boldsymbol{\nu}_{l,c2}$, $\forall\{c1,c2\} \in \{\{1,...,C1\},\{1,...,C2\}\}$.

Two measures $\varrho(\mathbf{y}_{l,\{c,k\}}, \mathbf{y}_{l,\{c1,k1\}}) = \|\mathbf{y}^-_{l,\{c,k\}} \odot \mathbf{y}^-_{l,\{c1,k1\}}\|_1 + \|\mathbf{y}^+_{l,\{c,k\}} \odot \mathbf{y}^+_{l,\{c1,k1\}}\|_1$ and $\varsigma(\mathbf{y}_{l,\{c,k\}}, \mathbf{y}_{l,\{c1,k1\}}) = \|\mathbf{y}_{l,\{c,k\}} \odot \mathbf{y}_{l,\{c1,k1\}}\|_2^2$ are used, where $\mathbf{y}_{l,\{c,k\}} = \mathbf{y}^+_{l,\{c,k\}} - \mathbf{y}^-_{l,\{c,k\}}$, $\mathbf{y}_{l,\{c1,k1\}} = \mathbf{y}^+_{l,\{c1,k1\}} - \mathbf{y}^-_{l,\{c1,k1\}}$, $\mathbf{y}^+_{l,\{c,k\}} = \max(\mathbf{y}_{l,\{c,k\}}, \mathbf{0})$ and $\mathbf{y}^-_{l,\{c,k\}} = \max(-\mathbf{y}_{l,\{c,k\}}, \mathbf{0})$. The measure $\varrho(\mathbf{y}_{l,\{c,k\}}, \mathbf{y}_{l,\{c1,k1\}})$ is related to similarity. Since, when $\mathbf{y}^T_{l,\{c,k\}}\mathbf{y}_{l,\{c1,k1\}}$ is considered, $\varrho(\mathbf{y}_{l,\{c,k\}}, \mathbf{y}_{l,\{c1,k1\}})$ captures contribution for the similarity, whereas $\|\mathbf{y}^+_{l,\{c,k\}} \odot \mathbf{y}^-_{l,\{c1,k1\}}\|_1 + \|\mathbf{y}^-_{l,\{c1,k1\}} \odot \mathbf{y}^+_{l,\{c1,k1\}}\|_1$ captures the contribution for the dissimilarity between the vectors $\mathbf{y}_{l,\{c,k\}}$ and $\mathbf{y}_{l,\{c1,k1\}}$. On the other hand, ς measures only the strength on the support intersection. The motivation is to use these measures and impose a discrimination constraint without any explicit assumption about the space/manifold in the transform domain.

A discriminative assignment w.r.t. to the parameters $\boldsymbol{\theta}_l$ that describe regions of similarity and dissimilarity in a collaborative way represents a trade-off between three elements: (i) similarity contribution, (ii) dissimilarity contribution and (iii) uncertainty corrective contribution w.r.t. a measure. A min-max functional $l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}})$ is one particular form that can be used to measures the score of this trade-off that we define it as follows:

$$l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}}) = \frac{1}{\beta_I} \min_{1\leq c1\leq C1} \max_{1\leq c2\leq C2} (\varrho(\mathbf{y}_{l,\{c,k\}}, \boldsymbol{\tau}_{l,c1}) + \varrho(\mathbf{y}_{l,\{c,k\}}, \boldsymbol{\nu}_{l,c2}) + \varsigma(\mathbf{y}_{l,\{c,k\}}, \boldsymbol{\tau}_{l,c1})), \tag{4}$$

where β_I is scaling parameter. By assumption $\boldsymbol{\tau}_{l,c1}$ and $\boldsymbol{\nu}_{l,c2}$ are spread far apart and cover the corresponding similarity and dissimilarity regions in the transform domain. Therefore, the min-max cost $l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}})$ ensures that $\mathbf{y}_{l,\{c,k\}}$ in the transform domain will be located at the point where (i) the similarity contribution w.r.t. $\boldsymbol{\tau}_{l,c1}$ is the smallest measured w.r.t. ϱ, (ii) the strength of the support intersection w.r.t. $\boldsymbol{\tau}_{l,c1}$ is the smallest measured w.r.t. ς, and (iii) the similarity contribution w.r.t. $\boldsymbol{\nu}_{l,c2}$ is the largest measured w.r.t. ϱ.

V. Joint Learning of Nonlinear Transforms with Targeted Self-Collaboration

In this section, first we explain the connections of the approximative log likelihood to an empirical risk and reveal the learning objective. then we presents the problem formulation.

A. Discriminating Self-Collaboration Likelihood as Empirical Risk

The learning goal is to estimate the parameters $\mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}$ and $\mathbf{A}$ that model the joint probability $p(\mathbf{x}_{c,k}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}, \mathbf{A})$ such that maximize the discrimination and collaboration specific log likelihood of the representations $\mathbf{y}_{l,\{c,k\}}$.

We take into account an approximation to the empirical expectation of the log likelihood of our model that factors into (i) the transform model probability $p(\mathbf{x}_{c,k}|\mathbf{Y}_{\{c,k\}}, \mathbf{A})$ and (ii) the discrimination probability prior $p(\boldsymbol{\theta}, \mathbf{Y}_{l,\{c,k\}})$, *i.e.*,

$$-\mathbb{E}[\log p(\mathbf{x}_{c,k}|\mathbf{Y}_{\{c,k\}}, \mathbf{A})] - \mathbb{E}[\log p(\boldsymbol{\theta}, \mathbf{Y}_{l,\{c,k\}})], \quad (5)$$

where we note that the explicit modeling of the parameters $\boldsymbol{\theta}_l$ results into clustering (or classification) based on maximum discrimination likelihood principle over a functional measure, and the estimation of $\mathbf{y}_{l,\{c,k\}}$ is in fact an assignment.

– *Empirical Risk:* Let $\mathbf{y}_{l,\{c,k\}}$ be given, we denote its cumulative cost w.r.t. the discrimination and collaboration corrective parameters as:

$$\begin{aligned}\xi_{l,\{c,k\}} =& \mathbf{t}_{l,\{c,k\}}^T|\mathbf{y}_{l,\{c,k\}}| + \mathbf{c}_{l,\{c,k\}}^T\mathbf{y}_{l,\{c,k\}} + \\ & \mathbf{n}_{l,\{c,k\}}^T(\mathbf{y}_{l,\{c,k\}} \odot \mathbf{y}_{l,\{c,k\}})\end{aligned} \quad (6)$$

where $\mathbf{t}_{l,\{c,k\}} = \max(\boldsymbol{\tau}_{l,c}, \mathbf{0}) \odot \text{sign}(\max(\mathbf{u}_{l,\{c,k\}}, \mathbf{0})) + \max(-\boldsymbol{\tau}_{l,c}, \mathbf{0}) \odot \text{sign}(\max(-\mathbf{u}_{l,\{c,k\}}, \mathbf{0})) - (\max(\boldsymbol{\nu}_{l,c}, \mathbf{0}) \odot \text{sign}(\max(\mathbf{u}_{l,\{c,k\}}, \mathbf{0})) + \max(-\boldsymbol{\nu}_{l,c}, \mathbf{0}) \odot \text{sign}(\max(-\mathbf{y}_{l,\{c,k\}}, \mathbf{0})))$ and $\mathbf{n}_{l,\{c,k\}} = \boldsymbol{\tau}_{l,c} \odot \boldsymbol{\tau}_{l,c}$. In fact, the empirical expectation of the cost (6), *i.e.*, $P_E : \frac{1}{CK}\sum_{c,k}\xi_{l,\{c,k\}} \simeq \mathbb{E}[\xi_{l,\{c,k\}}]$ represents the empirical risk for the proposed nonlinear transform model w.r.t. the used self-collaboration component and discrimination prior.

Therefore, we say that the learning objective is to estimate the model parameters that maximize a collaboration-corrective discriminative log likelihood (5). Or, in other words, during learning we target to estimate the model parameters that minimize the empirical risk (P_E).

B. The Problem Formulation

A joint maximization of $p(\mathbf{x}_{c,k}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}, \mathbf{A})$ over $\mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}$ and $\mathbf{A}$ is difficult. Instead, we consider minimizing $\mathbb{E}[-\log p(\mathbf{x}_{c,k}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}|\mathbf{A})p(\mathbf{A})]$, where $p(\mathbf{x}_{c,k}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}|\mathbf{A}) = p(\mathbf{x}_{c,k}|\mathbf{Y}_{\{c,k\}}, \mathbf{A})\prod_l p(\boldsymbol{\theta}_l|\mathbf{y}_{l,\{c,k\}})p(\mathbf{y}_{l,\{c,k\}})$. Moreover, given a data set $\mathbf{X}$, we minimize the unnormalized empirical approximation of the negative log likelihood for our model. The considered problem has the following form:

$$\begin{aligned}\min_{\boldsymbol{\Omega}} \sum_l \Bigg(& \frac{1}{2}\|\mathbf{A}_l\mathbf{X} - \mathbf{Y}_l\|_F^2 + \sum_{c,k}(l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}}) + \lambda_{l,1}\|\mathbf{y}_{l,\{c,k\}}\|_1) \\ & + \frac{1}{L}Tr\{(\mathbf{A}_l\mathbf{X} - \mathbf{Y}_l)^T\sum_{l1\neq l}(\mathbf{A}_{l1}\mathbf{X} - \mathbf{Y}_{l1})\} + \Omega(\mathbf{A}_l)\Bigg), \end{aligned} \quad (7)$$

where $\boldsymbol{\Omega} = \{\mathbf{Y}, \boldsymbol{\theta}, \mathbf{A}\}$, $\mathbf{Y} = [\mathbf{Y}_1, ..., \mathbf{Y}_L]$, $\mathbf{A} = [\mathbf{A}_1, ..., \mathbf{A}_L]$, and $\lambda_{l,1}$ is inversely proportional to the scaling parameter $\beta_{1,0}$.

– *Integrated Marginal Maximization*: We highlight that for our model, the solution to (7) is not equivalent to the maximum a posterior (MAP) solution[1], which would be difficult to compute, as it involves integration over $\mathbf{x}_{c,k}$, $\mathbf{y}_{l,\{c,k\}}$ and $\boldsymbol{\theta}$. Instead, we perform an integrated marginal minimization that is addressed with (7) and solved by iteratively marginally maximizing $p(\mathbf{x}_{c,k}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}, \mathbf{A})$ in $\mathbf{A}$, $\mathbf{Y}_{\{c,k\}}$ and $\boldsymbol{\theta}$. This is equivalent to 1) maximizing the conditional $p(\mathbf{x}_{c,k}|\mathbf{Y}_{\{c,k\}}, \mathbf{A})$ with prior $p(\mathbf{A}) = \prod_l p(\mathbf{A}_l)$ over $\mathbf{A}$, 2) maximizing the conditional $p(\mathbf{x}_{c,k}|\mathbf{Y}_{\{c,k\}}, \mathbf{A})$ with prior $p(\boldsymbol{\theta}_l|\mathbf{y}_{l,\{c,k\}})$ over $\mathbf{y}_{l,\{c,k\}}$ and 3) maximizing the conditional $p(\mathbf{y}_{l,\{c,k\}}|\boldsymbol{\theta}_l)$ over $\boldsymbol{\theta}_l$. The algorithm allows us to find a joint local maximum in $\{\mathbf{A}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}\}$ for $p(\mathbf{x}_{c,k}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}, \mathbf{A})$, such that the discrimination and collaboration specific prior probability is maximized (or in other words, the expected unnormalized negative log likelihood (5) is minimized).

In the following section, we present the learning algorithm through which we solve (7).

VI. The Learning Algorithm

As a solution to (7), we propose an iterative, alternating algorithm with three distinct stages: (i) representation $\mathbf{y}_{l,\{c,k\}}$ estimation with discriminative assignment, (ii) discrimination parameters $\boldsymbol{\theta}$ estimation and (iii) linear map $\mathbf{A}_l$ estimation. At the same time, we show that the problems at all stages have an exact or approximate closed-form solutions.

A. Discriminative Representation $\mathbf{y}_{l,\{c,k\}}$ Estimation

Given the available data samples $\mathbf{X}$ and the current estimate $\mathbf{A}_l$, the discriminative representation estimation problem per $\mathbf{Y}_l$ is decoupled and is formulated as: (P_{DRE}) : $\min_{\mathbf{Y}_l}\|\mathbf{A}_l\mathbf{X} - \mathbf{Y}_l\|_F^2 + \frac{1}{L}Tr\{\mathbf{Y}_l^T\sum_{l1\neq l}(\mathbf{Y}_{l1} - \mathbf{A}_{l1}\mathbf{X}))\} + \sum_{c,k}(l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}}) + \lambda_{l,1}\|\mathbf{y}_{l,\{c,k\}}\|_1)$.

We propose a solution for this stage that consists of two steps: (i) estimation of nonlinear transforms and (ii) nonlinear transform assignment based on the min-max functional discrimination score.

– *Nonlinear Transforms Estimation* Given all $\mathbf{Y}_l$ except $\mathbf{y}_{l,\{c,k\}}$, denote $\mathbf{c}_{l,\{c,k\}} = \frac{1}{2L}\sum_{l1}(\mathbf{y}_{l1,\{c,k\}} - \mathbf{A}_{l1}\mathbf{x}_{c,k})$, $\mathbf{b} = \mathbf{b}_{l,\{c,k\}} = \mathbf{A}_l\mathbf{x}_{c,k} + \mathbf{c}_{l,\{c,k\}}$ then for any $\mathbf{y} = \mathbf{y}_{l,\{c,k\}}$, problem (P_{DRE}) reduces to a *constrained projection* $(P_{DRE-R}) : \mathbf{y} = \arg\min_{\mathbf{y}}\|\mathbf{b} - \mathbf{y}\|_2^2 + l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}}) + \lambda_{l,1}\|\mathbf{y}_{l,\{c,k\}}\|_1$.

Assuming that $(\mathbf{v} - \mathbf{g})^T|\mathbf{y}| \geq 0$, per each pair $\{\boldsymbol{\tau}_{l,c1}, \boldsymbol{\nu}_{l,c2}\}$,$\{c1, c2\} \in \{\{1, ..., C1\} \times \{1, ..., C2\}\}$, (P_{DRE-R}) has a closed-form solution as:

$$\mathbf{y}|_{\{c1,c2\}} = \text{sign}(\mathbf{b}) \odot \max(|\mathbf{b}| - \mathbf{p}, \mathbf{0}) \oslash \mathbf{n}, \quad (8)$$

where $\mathbf{p} = \lambda_{l,0}(\mathbf{v} - \mathbf{g}) - \lambda_{l,1}\mathbf{1}$, $\mathbf{n} = (\mathbf{1} + 2\lambda_{l,0}\mathbf{s})$, $\mathbf{g} = \text{sign}(\max(\mathbf{q}, \mathbf{0})) \odot \mathbf{d}_1^+ + \text{sign}(\max(-\mathbf{q}, \mathbf{0})) \odot \mathbf{d}_1^-$, $\mathbf{v} = \text{sign}(\max(\mathbf{q}, \mathbf{0})) \odot \mathbf{d}_2^+ + \text{sign}(\max(-\mathbf{q}, \mathbf{0})) \odot \mathbf{d}_2^-$, $\mathbf{d}_1^+ = \boldsymbol{\tau}_{l,c1}^+$, $\mathbf{d}_1^- = \boldsymbol{\tau}_{l,c1}^-$, $\mathbf{d}_2^+ = \boldsymbol{\nu}_{l,c2}^+$, $\mathbf{d}_2^- = \boldsymbol{\nu}_{l,c2}^-$ and $\mathbf{s} = \boldsymbol{\tau}_{l,c1} \odot \boldsymbol{\tau}_{l,c1}$. (*the proof is given in Appendix A*).

– *Discriminative Assignment* This step consists of two parts.

[1]The MAP estimation problem for our model is identical to (7), but has additional terms that are related to the partition functions of $p(\mathbf{x}_{c,k}|\mathbf{Y}_{\{c,k\}}, \mathbf{A})$ and $p(\boldsymbol{\theta}, \mathbf{Y}_{\{c,k\}})$.

Part 1 Given the estimated $\mathbf{y}|_{\{c1,c2\}}$, $\{c1,c2\} \in \{1,...,C1\}\times\{1,...,C2\}$, the first part evaluates a score related to $l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}})$ as follows:

$$l_I : s_P(c1,c2) = \varrho(\mathbf{y}|_{\{c1,c2\}}, \boldsymbol{\tau}_{l,c1}) - \varrho(\mathbf{y}|_{\{c1,c2\}}, \boldsymbol{\nu}_{l,c2}) + \varsigma(\mathbf{y}|_{\{c1,c2\}}, \boldsymbol{\tau}_{l,c1}). \quad (9)$$

Part 2 Based on the score (9), the second part assigns $\mathbf{y}|_{\{c1,c2\}}$ to $\mathbf{y}_{l,\{c,k\}}$ using:

$$\{\widehat{c1},\widehat{c2}\} = \arg\min_{c1,c2} s_I(c1,c2), \qquad \mathbf{y}_{l,\{c,k\}} = \mathbf{y}|_{\{\widehat{c1},\widehat{c2}\}}. \quad (10)$$

Note that the maximum discrimination likelihood w.r.t. $l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}})$ is equivalent to computing a minimum score over s_I as in (10).

B. Parameters θ Update

Given the estimated representations $\mathbf{y}_{l,\{c,k\}}$, we update the parameters $\boldsymbol{\theta}_l, \forall l \in \{1,..,L\}$. Note that in the discriminative assignment (step (ii), parts 1 and 2), we have $(A_{ss}) : \{\mathbf{y}_{l,\{c,k\}} : \{\mathbf{y}|_{\{c1,c2\}}, \boldsymbol{\tau}_{l,c1}, \boldsymbol{\nu}_{l,c2}\}\}$, that is for each $\mathbf{y}_{l,\{c,k\}}$ the corresponding $\boldsymbol{\tau}_{l,c1}$ and $\boldsymbol{\nu}_{l,c2}$ are known. Therefore, $l_I(\boldsymbol{\theta}_l, \mathbf{y}_{l,\{c,k\}})$ is not evaluated at this stage, instead we use (A_{ss}) for the update of the parameters $\boldsymbol{\theta}_l$.

– *Update per single* $\boldsymbol{\tau}_{l,c1}$: Using (A_{ss}), we formulate the problem associated to the update of $\boldsymbol{\tau}_{l,c1}$ as follows:

$$\boldsymbol{\tau}_{l,c1} = \arg\min_{\boldsymbol{\tau}_{l,c1}} \frac{1}{2}\|\boldsymbol{\tau}_{l,c1}^{t-1} - \boldsymbol{\tau}_{l,c1}\|_2^2 + \lambda_{l,0}\sum_{c1}(\varsigma(\mathbf{y}|_{\{c1,c2\}}, \boldsymbol{\tau}_{l,c1}) + \varrho(\mathbf{y}|_{\{c1,c2\}}, \boldsymbol{\tau}_{l,c1})), \quad (11)$$

where $\boldsymbol{\tau}_{l,c1}^{t-1}$ and $\boldsymbol{\tau}_{l,c1}$ are the parameters at iteration $t-1$ and t, and $\lambda_{l,0}$ is inversely proportional to the scaling parameter β_I. The solution to (11) is similar to the solution (8), the difference is that in the solution for (11) the respective thresholding is different and there is no normalization vectors (the proof is given in *Appendix B.1*).

– *Update per single* $\boldsymbol{\nu}_{l,c2}$: Similarly, we use (A_{ss}) and formulate the problem related to update of $\boldsymbol{\nu}_{l,c2}$ as $\boldsymbol{\nu}_{l,c2} = \arg\min_{\boldsymbol{\nu}_{l,c2}} \frac{1}{2}\|\boldsymbol{\nu}_{l,c2}^{t-1} - \boldsymbol{\nu}_{l,c2}\|_2^2 + \lambda_{l,0}\sum_{c2}\varsigma(\mathbf{y}|_{\{c1,c2\}}, \boldsymbol{\nu}_{l,c2})$, where $\boldsymbol{\nu}_{c2}^{t-1}$ and $\boldsymbol{\nu}_{c2}$ denotes the parameters at iteration $t-1$ and t. Again, the solution here is similar to the solution (8) with the difference that in the solution for $\boldsymbol{\nu}_{l,c2}$ there is no thresholding and the normalization vector is different (the proof is given in *Appendix B.2*).

C. Linear Map $\mathbf{A}_l$ Estimation

Given the data samples $\mathbf{X}$, all $\mathbf{Y} = [\mathbf{Y}_1,...,\mathbf{Y}_L]$, and all $\mathbf{A}$ except $\mathbf{A}_l$, denote $\mathbf{W}_l = \mathbf{Y}_l - \sum_{l1}(\mathbf{A}_{l1}\mathbf{X} - \mathbf{Y}_{l1})$ then the problem related to the estimation of the linear map $\mathbf{A}_l$, reduces to $\min_{\mathbf{A}_l} \frac{1}{2}\|\mathbf{A}_l\mathbf{X} - \mathbf{W}_l\|_2^2 + \frac{\lambda_{l,2}}{2}\|\mathbf{A}_l\|_F^2 + \frac{\lambda_{l,3}}{2}\|\mathbf{A}_l\mathbf{A}_l^T - \mathbf{I}\|_F^2 - \lambda_{l,4}\log|\det\mathbf{A}_l^T\mathbf{A}_l|$, where $\{\lambda_{l,2}, \lambda_{l,3}, \lambda_{l,4}\}$ are inversely proportional to the scaling parameters $\{\beta_{l,3}, \beta_{l,4}, \beta_{l,5}\}$ and we use approximate closed-form solution as proposed in [17].

VII. Evaluation of the Proposed Approach

In this section we evaluate the algorithm properties, the discriminative quality, and the recognition accuracy.

	AR	YALE B	COIL20	NORB
$\frac{1}{L}\sum_l \mu(\mathbf{A}_l)$	2.1e-4	1e-4	1.9e-4	3.1e-4
$\frac{1}{L}\sum_l C_n(\mathbf{A}_l)$	16.1	26.3	18	19.1

TABLE I
THE CUMULATIVE EXPECTED MUTUAL COHERENCE $\frac{1}{L}\sum_l \mu(\mathbf{A}_l)$ AND THE CUMULATIVE CONDITIONING NUMBER $\frac{1}{L}\sum_l C_n(\mathbf{A}_l)$ FOR THE LINEAR MAPS $\mathbf{A}_l, l \in \{1,...,6\}$ WITH DIMENSIONS $6570 \times N$, WHERE N IS THE DIMENSIONALITY OF THE INPUT DATA

	AR	YALE B	COIL20	NORB
learning time $[h]$	$L\times$ .212	$L\times$ .301	$L\times$ .111	$L\times$ .261

TABLE II
THE LEARNING TIME IN HOURS ON THE DATABASES AR, YALE B, COIL20 AND NORB USING OUR MODEL WITH DIMENSION $M = 6570$, NUMBER OF SELF-COLLABORATION COMPONENTS $L = 9$, AND DIMENSION PER SELF-COLLABORATION COMPONENT $M/L = 730$.

A. Quantifying a Discrimination Quality

To quantify the discriminative properties of a dataset under a transform, we introduces a measure about the discrimination properties of a dataset.

The discriminative properties of a dataset under a transform with parameter set $\mathcal{P}_t = \{\mathbf{A} = [\mathbf{A}_1,...,\mathbf{A}_L]^T \in \Re^{M\times N}, \tau\mathbf{1} \in \Re^M\}$ are defined using the relations between two concentrations $D_{\ell_1}^{\mathcal{P}_t}(\mathbf{X})$ and $D_{\ell_1,c}^{\mathcal{P}_t}(\mathbf{X})$. The first is the expected similarity of all $\mathbf{u}_{c,k} = [\mathbf{y}_{1,\{c,k\}}^T,...,\mathbf{y}_{L,\{c,k\}}^T]^T$ across all the transform representations $\mathbf{Y}_c$ that come from the different classes $c1 \neq c$, *i.e.*, $D_{\ell_1}^{\mathcal{P}}(\mathbf{X}) = \sum_c\sum_{c1\neq c}\sum_{k=1}^K\sum_{k1\neq k}(\|\mathbf{u}_{c,k}^+ \odot \mathbf{u}_{c1,k1}^+\|_1 + \|\mathbf{u}_{c,k}^- \odot \mathbf{u}_{c1,k1}^-\|_1)$. The second is the expected similarity using the positive and negative components of all $\mathbf{u}_{c,k} = [\mathbf{y}_{1,\{c,k\}}^T,...,\mathbf{y}_{L,\{c,k\}}^T]^T$ across all the transform representations $\mathbf{Y}_c$ that come from the same classes c, *i.e.*, $D_{\ell_1,c}^{\mathcal{P}}(\mathbf{X}) = \sum_c\sum_{k=1}^K\sum_{k1\neq k}(\|\mathbf{u}_{c,k}^+ \odot \mathbf{u}_{c,k1}^+\|_1 + \|\mathbf{u}_{c,k}^- \odot \mathbf{u}_{c,k1}^-\|_1)$. The *discrimination power* for any pair of labels and dataset $\mathbf{X} \in \Re^{M\times CK}$ under a transform with parameter set $\mathcal{P}_t = \{\mathbf{A} \in \Re^{M\times N}, \tau\mathbf{1} \in \Re^M\}$ is defined as:

$$\mathcal{I}^t = \log(D_{\ell_1,c}^{\mathcal{P}_t}(\mathbf{X})) - \log(D_{\ell_1}^{\mathcal{P}_t}(\mathbf{X}) + \epsilon), \quad (12)$$

where $\epsilon > 0$ is small constant. By letting $\mathbf{A} = \mathbf{I}$ and $\tau = 0$ in $\mathcal{P}_t$, (12) allows to measure the discrimination power for any given data set $\mathbf{X}$. The advantage of this measure is that it logarithmically signifies the difference between $D_{\ell_1,c}^{\mathcal{P}_t}(\mathbf{X})$ and $D_{\ell_1}^{\mathcal{P}_t}(\mathbf{X})$. The numerical evaluation shows that this measure is related to the recognition capabilities and that gives insight into the learning dynamics of the proposed algorithm.

B. Data Sets and Algorithms Set Up

The used datasets are AR [23], Extended YALE B [7], COIL20 [25], NORB [20], MNIST [19], F-MNIST [41] and SVHN [26]. All the images from the respective datasets were downscaled to resolutions 32×28, 21×21, 20×25, 24×24, 28×28 and 28×28, and are normalized to unit variance.

An on-line variant is used for the update of $\mathbf{A}$ w.r.t. a subset of the available training set. It has the following form $\mathbf{A}^{t+1} = \mathbf{A}^t - \rho(\mathbf{A}^t - \hat{\mathbf{A}})$ where $\hat{\mathbf{A}}$ and $\mathbf{A}^t$ are the the solutions in the transform update step at iterations $t+1$ and t, which is equivalent to having the additional constraint $\|\mathbf{A}^t - \hat{\mathbf{A}}\|_F^2$ in the related problem. The used batch size is equal to $87\%, 85\%, 90\%, 87\%, 5\%, 5\%$ and 5% of the total amount

	AR	YALE B	COIL20	NORB
$\mathcal{I}^o$	2.13	1.45	1.18	0.41
$\mathcal{I}^R$	2.41	1.66	1.61	0.40
$\mathcal{I}^S$	2.71	1.76	1.92	0.40
$\mathcal{I}^*$	3.04	2.14	2.63	0.42

TABLE III
THE DISCRIMINATION POWER IN THE ORIGINAL DOMAIN, AFTER RANDOM TRANSFORM, AFTER LEARNED SPARSIFYING TRANSFORM AND AFTER LEARNED SELF-COLLABORATING TARGET SPECIFIC NONLINEAR TRANSFORM WITH DIMENSION $M = 6570$.

	AR	YALE B	COIL20	NORB
original domain [%]	96.1	95.4	96.8	97
proposed [%]	97.1	97.1	97.8	96.8

TABLE IV
THE RECOGNITION RESULTS ON THE DATABASES AR, YALE B, COIL20 AND NORB, USING K-NN ON THE SPARSE REPRESENTATIONS USING OUR MODEL WITH DIMENSION $M = 6570$.

of the available training data from the respective datasets AR, Extended YALE B, COIL20, NORB , MNIST, F-MNIST and SVHN.

C. Numerical Experiments

Sparsifying Nonlinear Transform (sNT) Representation In the numerical experiments we learn our model using the proposed algorithm. Next, we construct a sparsifying nonlinear transform (sNT) representation using our learned model by (i) computing sparsifying transforms as $\mathbf{u}_{l,\{c,k\}} = \text{sign}(\mathbf{A}_l \mathbf{x}_{c,k}) \odot \max(|\mathbf{A}_l \mathbf{x}_{c,k}| - \tau \mathbf{1}, \mathbf{0})$ and (ii) concatenating them as $\mathbf{u}_{c,k} = [\mathbf{u}_{1,\{c,k\}}^T, ..., \mathbf{u}_{L,\{c,k\}}^T]^T$.

The numerical experiments are preformed using the sNT representations and consist of three parts:

– *Model Properties:* In the first series of the experiments, we evaluate the cumulative expected mutual coherence $\sum_l \mu(\mathbf{A}_l)$ (where $\mu(\mathbf{A}_l)$ is computed as in [17]), the cumulative conditioning number $\sum_l C_n(\mathbf{A}_l)$, $C_n(\mathbf{A}) = \frac{\sigma_{\max}}{\sigma_{\min}}$ ($\sigma_{\max}$ and $\sigma_{\min}$ are the maximal and minimal singular values of $\mathbf{A}$, respectively) and the computational efficiency as run time $t[h]$.

– *Discrimination Power of the sNT Representation:* A comparison is presented between the discrimination power $\mathcal{I}$ under different transforms. The $\mathcal{I}$ is estimated in the original domain, after transform by a Gaussian random matrix and after a learned nonlinear transform having transform dimension $M = 6570$ without and with discriminative prior, denoted as $\mathcal{I}^o$, $\mathcal{I}^R$, $\mathcal{I}^S$ and $\mathcal{I}^*$, respectively.

– *Recognition Accuracy Comparison Proposed vs State-Of-The-Art:* The third part evaluates the discrimination power and the recognition accuracy using the representations from our model as features and compares it to several state-of-the-art (sota) methods, including 1) supervised dictionary learning methods [28], [43], [38] and [37], 2) unsupervised feature learning methods [26], [13], [27], 3) different classifiers [40] and 4) deep neural networks [9], [22], [21], [34].

This comparison considers a setup where the used data are divided into a training and test set. The learning is performed on the training set and the evaluation is performed on the test set. The training sNT representations are used to estimate the SVM parameters and the recognition is performed using the learned SVM on the test sNT representations.

	YALE B $\mathcal{I}$	MNIST $\mathcal{I}$		YALE B Acc. [%]		MNIST Acc. [%]
$dlsi$ [28]	0.71	0.67		96.5		98.74
$fddl$ [43]	0.87	0.63		97.5		96.31
$copar$ [38]	0.57	0.54		98.3		96.41
$lrsdl$ [37]	0.42	0.40		98.7		–
$*$	0.90	0.81	k-nn	97.1	k-nn	97.32
$*$	0.90	0.81	l-svm	98.8	l-svm	98.45
a)			b)		c)	

TABLE V
A) THE DISCRIMINATION POWER FOR THE METHODS $dlsi$[28], $fddl$ [43], $copar$ [38] AND $lrsdl$ [37] AND THE PROPOSED METHOD $*$, B), C) THE RECOGNITION RESULTS ON THE EXTENDED YALE B AND MNIST

D. Results

– *Model Properties:* The cumulative conditioning number $\frac{1}{L}\sum_l C_n(\mathbf{A}_l)$ and the cumulative expected coherence $\frac{1}{L}\sum_l \mu(\mathbf{A}_l)$ for the learned transforms using the databases AR, YALE B, COIL20 are shown in Table I. All linear maps per all the databases have good conditioning numbers and low expected coherence. This confirms the effectiveness of the conditioning and the coherence constraints. The running time $t[h]$, measured in hours for learning the model parameters with $M = 6570$ are shown in Table II. The learned transforms for all the datasets have relatively low execution time, despite the very high transform dimension.

– *Discrimination Quality:* The results are shown in Table III. The discrimination power $\mathcal{I}^*$ is significantly increased in the transform domain compared to the one in the original domain $\mathcal{I}^O$ and is higher than $\mathcal{I}^R$ and $\mathcal{I}^S$.

– *Unsupervised Classification Recognition Accuracy:* Table IV shows the recognition results on the databases AR, YALE B, COIL20 and NORB using k-nn as a classifier, where we compare the results w.r.t. to the baseline, that is a k-nn in the original domain and we see improvements. Where as for the results shown on Tables V and VI we see comparable results.

– *Proposed vs State-Of-The-Art* Considering the evaluation of the discrimination power, in all the algorithms, the dictionary size (transform dimension M) is set to be equal to $\{150, 300\}$ for the used databases, respectively. The discrimination power is compared with the methods dlsi, fddl, copar and lrsdr. The results are shown in Table V. The discrimination power of the sNT representation is higher than the discrimination power of the comparing methods.

At the databases YALE B and MNIST the recognition accuracy is also comparable and higher, respectively, w.r.t. the state-of-the-art DDL methods. The results shown in Table VI demonstrate improvements and competitive performance w.r.t. the comparing methods for unsupervised feature learning.

Considering the comparison w.r.t. deep neural networks, [9], [34] and [21] achieve highest accuracy on MNIST, F-MNIST and SVHN. We highlight that although we learn a target specific self-collaboration model with discriminative and sparsity priors, during testing we use a simple sNT representation. Whereas in most of the deep neural networks [13], [22], [27], [9], [34], [26] and [21], the modeling by a multi-layer architecture, the local image content relations and/or aggregation and nonlinearities were taken into account.

MNIST		F-MNIST		SVHN	
Method	Acc.	Method	Acc.	Method	Acc.
lif-cnn [13]	98.37	log-reg [40]	84.00	ssae [26]	89.70
s-cw-a [22]	98.62	rf-c [40]	87.70	c-km [26]	90.60
reg-l [27]	99.08	svc [40]	89.98	s-cw-a [22]	93.10
f-max [9]	99.65	cnn [34]	92.10	tma [21]	98.31
$*$ k-nn	97.11	$*$ k-nn	88.10	$*$ k-nn	86.41
$*$ l-svm	99.10	$*$ l-svm	92.22	$*$ l-svm	90.28

TABLE VI
RECOGNITION ACCURACY COMPARISON BETWEEN SOTA AND 1) K NEAREST NEIGHBOR (k-nn) SEARCH AND 2) LINEAR SVM [12] (l-svm) THAT USE THE SPARSIFYING NONLINEAR TRANSFORM (SNT) REPRESENTATIONS FROM OUR MODEL ON EXTRACTED HOG [5] IMAGE FEATURES. WE USE OUR ALGORITHM TO LEARN THE MODEL ON THE HOG FEATURES. THEN WE GET THE SNT REPRESENTATIONS WITH DIMENSIONALITY 9800 FOR THE RESPECTIVE TRAINING AND TEST SETS. CONSIDERING THE OBTAINED RESULT FOR DATABASE SVHN, WE NOTE THAT THE UNLABELED TRAINING DATA FROM THE RESPECTIVE DATABASE WAS NOT USED DURING THE LEARNING OF THE CORRESPONDING MODEL.

VIII. CONCLUSION

This paper introduced a novel collaboration structured model with minimum information loss, collaboration corrective and discriminative priors for joint learning of multiple nonlinear transforms. The model parameters were learned by addressing an integrated marginal maximization that corresponds to minimizing an unnormalized empirical log likelihood of the model. An efficient solution was proposed by an iterative, coordinate descend algorithm.

The preliminary results w.r.t. the introduced measure and the recognition accuracy on the used databases showed promising performance and advantages w.r.t. state-of-the-art methods. A study on the recognition capabilities for other databases, together with a study on a deep architecture where per single layer we have nonlinear transform are left for future work.

APPENDIX A

Let $\mathbf{y} = \mathbf{y}_{l,\{c,k\}}$, $\mathbf{q} = \mathbf{A}_l\mathbf{x}_{c,k} + \mathbf{c}_{l,\{c,k\}}$ and $\lambda_0 = \lambda_{l,0}$. Denote $\mathbf{d}_d = \boldsymbol{\tau}_{l,c1}$, $\mathbf{d}_s = \boldsymbol{\nu}_{c2}$ and $\mathbf{s} = \boldsymbol{\tau}_{l,c1} \odot \boldsymbol{\tau}_{l,c1}$, $\mathbf{g}_d = (\text{sign}(\mathbf{q}^+) \odot \mathbf{d}_d^+ + \text{sign}(-\mathbf{q}^-) \odot \mathbf{d}_d^-)$, $\mathbf{g}_s = (\text{sign}(\mathbf{q}^+) \odot \mathbf{d}_s^+ + \text{sign}(-\mathbf{q}^-) \odot \mathbf{d}_s^-)$ and note that per pair $\{\boldsymbol{\tau}_{l,c1}, \boldsymbol{\nu}_{l,c2}\}$ we have the following problem:

$$\min_{\mathbf{y}} \frac{1}{2}\|\mathbf{q}-\mathbf{y}\|_2^2 + \lambda_0\big(\mathbf{g}_d^T|\mathbf{y}| - \mathbf{g}_s^T|\mathbf{y}|\big) + \mathbf{s}^T(\mathbf{y}\odot\mathbf{y}) + \lambda_1\mathbf{1}^T|\mathbf{y}|, \quad (13)$$

that represents a *projection problem with linear, quadratic and sparsity constraints*.

First Order Derivative The first order derivative w.r.t. $\mathbf{y}$ is:

$$\begin{aligned}\mathbf{y} - \mathbf{q} + \lambda_0\left(\mathbf{g}_d \odot \text{sign}(\mathbf{y}) - \mathbf{g}_s \odot \text{sign}(\mathbf{y})\right) + \\ \lambda_0\mathbf{y} \odot \mathbf{s} + \lambda_1\text{sign}(\mathbf{y}) = \mathbf{0},\end{aligned} \quad (14)$$

let $\mathbf{y} = |\mathbf{y}| \odot \text{sign}(\mathbf{y})$, $\mathbf{q} = |\mathbf{q}| \odot \text{sign}(\mathbf{q})$ and $\mathbf{k} = (\mathbf{1} + \lambda_0\mathbf{s})$ and assuming that $\text{sign}(\mathbf{y}) = \text{sign}(\mathbf{q})$, then we have $|\mathbf{y}| \odot \mathbf{k} - |\mathbf{q}| + \lambda_0\left(\mathbf{g}_d - \mathbf{g}_s\right) + \lambda_1\mathbf{1} = \mathbf{0}$. Hadamard divide by left with $\mathbf{k}$ we have that:

$$|\mathbf{y}| - |\mathbf{q}| \oslash \mathbf{k} + \lambda_0\left(\mathbf{g}_d - \mathbf{g}_s\right) \oslash \mathbf{k} + \lambda_1\mathbf{1} \oslash \mathbf{k} = \mathbf{0}, \quad (15)$$

since the magnitude might be only positive we have that $|\mathbf{y}| = \max(|\mathbf{q}| \oslash \mathbf{k} - \lambda_0\left(\mathbf{g}_d - \mathbf{g}_s\right) \oslash \mathbf{k} - \lambda_1\mathbf{1} \oslash \mathbf{k}, \mathbf{0})$.

Closed-Form Solution Assuming $\mathbf{k}^T(\mathbf{y} \odot \mathbf{y}) \geq 0$ and $\left(\mathbf{g}_d^T|\mathbf{y}| - \mathbf{g}_s^T|\mathbf{y}|\right) \geq 0$ then the closed-form solution is:

$$\mathbf{y} = \text{sign}\left(\mathbf{q}\right) \odot \max\left(|\mathbf{q}| - \lambda_0\left(\mathbf{g}_d - \mathbf{g}_s\right) - \lambda_1\mathbf{1}, \mathbf{0}\right) \oslash \mathbf{k} \ \square \quad (16)$$

APPENDIX B

B.1 Let all the variables be fixed and denote $\lambda_{l,0} = \lambda_0$, then per $\boldsymbol{\tau}_{l,c1}$ we have the following problem:

$$\min_{\boldsymbol{\tau}_{l,c1}} \frac{1}{2}\|\boldsymbol{\tau}_{l,c1}^{t-1} - \boldsymbol{\tau}_{l,c1}\|_2^2 + \lambda_0 \sum_{c1} r(c1), \quad (17)$$

where $r(c1) = \varrho(\mathbf{y}|_{\{c1,c2\}}, \boldsymbol{\tau}_{l,c1}) + \varsigma(\mathbf{y}|_{\{c1,c2\}}, \boldsymbol{\tau}_{l,c1})$. Let $\mathbf{y} = \boldsymbol{\tau}_{l,c1}$ and $\mathbf{q} = \boldsymbol{\tau}_{l,c1}^{t-1}$, denote $\mathbf{g}_d = (\text{sign}(\mathbf{q}^+) \odot \sum_{c1}\mathbf{y}|_{\{c1,c2\}}^+ + \text{sign}(-\mathbf{q}^-) \odot \sum_{c1}\mathbf{y}|_{\{c1,c2\}}^-)$ and $\mathbf{s} = \sum_{c1}\mathbf{y}|_{\{c1,c2\}} \odot \mathbf{y}|_{\{c1,c2\}}$, then we have the following problem:

$$\min_{\mathbf{y}} \frac{1}{2}\|\mathbf{q} - \mathbf{y}\|_2^2 + \lambda_0\mathbf{g}_d^T|\mathbf{y}| + \mathbf{s}^T(\mathbf{y} \odot \mathbf{y}), \quad (18)$$

that is essentially equivalent to problem (13), but, without the terms $-\lambda_0\mathbf{g}_s^T|\mathbf{y}|$ and $\lambda_1\mathbf{1}^T|\mathbf{y}|$.

Closed-Form Solution Assuming $\mathbf{k}^T(\mathbf{y} \odot \mathbf{y}) \geq 0$, then the closed-form solution is:

$$\mathbf{y} = \text{sign}\left(\mathbf{q}\right) \odot \max\left(|\mathbf{q}| - \lambda_0\mathbf{g}_d, \mathbf{0}\right) \oslash \mathbf{k}, \quad (19)$$

where similarly as in *Appendix A*, $\mathbf{k} = \mathbf{1} + \lambda_0\mathbf{s}$ $\square$

B.2 Let all the variables be fixed and denote $\lambda_{l,0} = \lambda_0$, then per $\boldsymbol{\nu}_{l,c2}$ we have the following problem:

$$\min_{\boldsymbol{\nu}_{l,c2}} \frac{1}{2}\|\boldsymbol{\nu}_{l,c2}^{t-1} - \boldsymbol{\nu}_{l,c2}\|_2^2 + \lambda_0 \sum_{c2} r(c2), \quad (20)$$

where $r(c2) = \varrho(\mathbf{y}|_{\{c1,c2\}}, \boldsymbol{\nu}_{l,c2})$. Let $\mathbf{y} = \boldsymbol{\nu}_{l,c2}$ and $\mathbf{q} = \boldsymbol{\nu}_{l,c2}^{t-1}$, denote $\mathbf{g}_s = (\text{sign}(\mathbf{q}^+) \odot \sum_{c2}\mathbf{y}|_{\{c1,c2\}}^+ + \text{sign}(-\mathbf{q}^-) \odot \sum_{c2}\mathbf{y}|_{\{c1,c2\}}^-)$ then we have the following problem:

$$\min_{\mathbf{y}} \frac{1}{2}\|\mathbf{q} - \mathbf{y}\|_2^2 + \lambda_0\mathbf{g}_d^T|\mathbf{y}|, \quad (21)$$

that is essentially equivalent to problem (13), but, without the terms $-\lambda_0\mathbf{g}_d^T|\mathbf{y}|$, $\lambda_1\mathbf{1}^T|\mathbf{y}|$ and $\lambda_0\mathbf{s}^T(\mathbf{y} \odot \mathbf{y})$.

Closed-Form Solution The closed-form solution is:

$$\mathbf{y} = \text{sign}\left(\mathbf{q}\right) \odot \max\left(|\mathbf{q}| - \lambda_0\mathbf{g}_s, \mathbf{0}\right) \ \square \quad (22)$$

REFERENCES

[1] Francis R Bach and Zaïd Harchaoui. Diffrac: a discriminative and flexible framework for clustering. In *NIPS*, pages 49–56, 2008.

[2] Francis R. Bach and Michael I. Jordan. Learning spectral clustering, with application to speech separation. *JMLR*, 7:1963–2001, 2006.

[3] Pierre Baldi. Autoencoders, unsupervised learning and deep architectures. In *UTLM*, pages 37–50, 2011.

[4] Richard G. Baraniuk, Volkan Cevher, Marco F. Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Trans. Inf. Theor.*, 56(4):1982–2001, April 2010.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. CVPR, 2005.

[6] Michal Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse Problems*, 23 (3):947–968, June 2007.

[7] 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 TPAMI*, pages 643–660, 2001.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[9] Benjamin Graham. Fractional max-pooling. *CoRR*, 2014.

[10] S. Hawe, M. Kleinsteuber, and K. Diepold. Analysis operator learning and its application to image reconstruction. *IEEE Trans. Image Processing*, 22(6):2138–2150, 2013.

[11] Geoffrey E. Hinton. Boltzmann machines. In *Encyclopedia of Machine Learning and Data Mining*. 2017.

[12] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Trans. on Neural Networks*, 2:415–425, 2002.

[13] Eric Hunsberger and Chris Eliasmith. Spiking deep networks with LIF neurons. 2015.

[14] Sofia Karygianni and Pascal Frossard. Structured sparse coding for image denoising or pattern detection. In *ICASSP, Florence, Italy*, 2014.

[15] Byung Soo Kim, Jae Young Park, Anush Mohan, Anna Gilbert, and Silvio Savarese. Hierarchical classification of images by sparse approximation. In *BMVC*, 2011.

[16] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems, 2009.

[17] D. Kostadinov, S. Volshinovsky, and Sohrab Ferdowsi. Learning non-structured, overcomplete and sparsifying transform. In *SPARS, Lisbon, Portugal*, June 2017.

[18] Andreas Krause, Pietro Perona, and Ryan G Gomes. Discriminative clustering by regularized information maximization. In *NISP*, 2010.

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

[20] Yann LeCun, Fu Jie Huang, and Léon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *CVPR*, pages 97–104, 2004.

[21] Chen-Yu Lee, Patrick W. Gallagher, and Zhuowen Tu. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In *19th ICAIS*, 2016.

[22] Alireza Makhzani and Brendan J Frey. Winner-take-all autoencoders. In *NIPS*. 2015.

[23] A. Martínez and R. Benavente. The ar face database. Technical report, Computer Vision Center, 1998.

[24] G. J. McLachlan and D. Peel. *Finite mixture models*. Wiley Series in Probability and Statistics, New York, 2000.

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

[26] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Wor. UFL*, 2011.

[27] Priyadarshini Panda and Kaushik Roy. Unsupervised regenerative learning of hierarchical features in spiking deep networks for object recognition. *CoRR*, 2016.

[28] I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *CVPR*, 2010.

[29] Saiprasad Ravishankar and Yoram Bresler. Learning sparsifying transforms for image processing. In *IEEE ICIP, Orlando, FL, USA*, pages 681–684, 2012.

[30] Saiprasad Ravishankar and Yoram Bresler. Closed-form solutions within sparsifying transform learning. In *IEEE ICASSP, Vancouver, BC, Canada*, pages 5378–5382, 2013.

[31] R. Rubinstein and M. Elad. Dictionary learning for analysis-synthesis thresholding. *IEEE Trans. Signal Processing*, 62(22):5962–5972, 2014.

[32] R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[33] R. Rubinstein, T. Peleg, and M. Elad. Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model. *IEEE TSP*, pages 661–677, 2013.

[34] Maheshkumar H. Kolekar Shobhit Bhatnagar, Deepanway Ghosal. Classification of fashion article images using convolutional neural networks. In *IEEE ICIIP*, 2017.

[35] Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE TNN*, pages 586–600, 2000.

[36] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 2010.

[37] T. H. Vu and V. Monga. Fast low-rank shared dictionary learning for image classification. *CoRR*, abs/1610.08606, 2016. URL http://arxiv.org/abs/1610.08606.

[38] 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 *IEEE ISBI*, 2015.

[39] Anja Wille and Peter Buhlmann. Low-order conditional independence graphs for inferring genetic networks. In *Statistical Applications in Genetics and Molecular Biology, 5.1*, pages 121–136. DE GRYTIER, 2018.

[40] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, 2017.

[41] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

[42] Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. In *Advances in neural information processing sys.*, pages 1537–1544, 2005.

[43] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *ICCV*, pages 543–550, 2011.