Supervised Joint Nonlinear Transform Learning with Discriminative-Ambiguous Prior for Generic Privacy-Preserved Features

Dimche Kostadinov, Behrooz Razeghi, Shideh Rezaeifar and Slava Voloshynovskiy
Stochastic Information Processing Group
Department of Computer Science, University of Geneva, Switzerland
{dimche.kostadinov, behrooz.razeghi, shideh.rezaeifar, svolos}@unige.ch

***Abstract*—In this paper, we explicitly model a discriminative-ambiguous setup by two jointly learned parametric nonlinear transforms. The idea is to use one nonlinear transform for ambiguization and the other one for discrimination, and also to address a privacy-utility setup that is conditioned on ambiguization and discrimination priors, respectively, together with minimum information loss prior. The generic coupled representation is composed by linear combination using the two nonlinear transforms. The model parameters are learned by minimizing the empirical log likelihood of the model, where we propose an efficient solution using block coordinate descend alternating algorithm. The proposed model has low computational complexity and high recognition accuracy for the authorized parties while having low recognition accuracy for the unauthorized parties. We validate the potential of the proposed approach by numerical experiments.**

***Index Terms*—nonlinear transform learning, privacy, utility, discrimination, ambiguization.**

I. Introduction

In the past few decades, the privacy-preserving data release and data mining become growing concerns due to the massive increase in personal information stored in the electronic data sources. The objective of privacy-preserving data release is to provide the useful data (utility) for a desired application while simultaneously protecting personal information (privacy). Several research communities proposed different definitions for privacy and utility, which are motivated by the desired functionality of applications. A key theoretical, machine learning and statistical challenge is to understand when the goals of utility and privacy can be efficiently achieved simultaneously [1]–[3]. In general, the privacy-utility goal depends on the data statistics, probabilistic model, adversarial strategy and objectives, used privacy and utility measures and also on the problem formulation.

There is a rich literature of prior work on privacy protection techniques. In summary, the common key approaches used to protect privacy can be categorized as: 1) randomization methods, which are mostly based on data perturbation techniques, 2) data swapping techniques, 3) secure multi-party computation techniques, and 4) distributed data mining techniques.

Our work is closely related to [4], [5], where a privacy-preserving identification scheme based on sparsifying transform learning with ambiguization was proposed. The idea in [4] was to obtain an information-preserving sparse ternary representation of the data, while ensuring privacy. The process is as follows: (i) a sparsifying transform model with a square orthonormal model matrix is learned, (ii) next using the learned model the data are sparsified and ternarized, (iii) then a perturbation noise is added to the zero components (complementary support) of the sparsifying transform representation, which was the final public representation (protected template).

The main open issue with the approach [4] is the optimality of the transform from two different perspectives. The first issue is that all the steps were computed independently and were not addressed in the learning problem, that is the sparsifying model is learned in step (i) independently from the steps (ii) and (iii), where a nonlinearity by ternary quantization is used and the perturbation noise is added and . The second issue is that even with a sub-optimal approach the privacy-utility setup is only addressed w.r.t. the quality of reconstruction from the transform to original domain that represents the utility. In addition, the method only learns a sparsifying transform and no additional prior for discrimination or distinctiveness between the transform data was included.

Contributions: Towards generic privacy-preserved features described through learned nonlinear transforms, we propose a novel model for learning nonlinear transforms, which are parameterized by linear maps and element-wise nonlinearities. The model entails minimum information loss, supervised discrimination and ambiguization priors together with a conditional privacy-utility prior, which is described over the nonlinear transform representations. The model characterizes a privacy-utility trade-off w.r.t. a notion for discrimination and ambiguization in a principle manner. The corresponding problem formulation allows achieving an extreme point through the joint learning of the two nonlinear transforms with priors by reducing or extending dimensionality while preserving discrimination and imposing ambiguization.

The proposed model can be incorporated in various applications such as privacy-preserving identification and recognition schemes [4], [6], [7], privacy-preserving classification and clustering schemes [8], and differentiate-private learning techniques [1], [9], [10], etc.

B. Razeghi has been supported by the ERA-Net project ID_IoT No 20CH21_167534 and S. Rezaeifar by the SNF project No 200021_182063.

While there are many interesting connections to the differentially privacy techniques [9], [11] and information-theoretic approaches [12]–[14], in this paper the main focus is on presenting the core idea in light of our model and at the same time demonstrating the potential and advantages by numerical evaluation. In the following subsection, we introduce the model and overview the learning strategy.

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 transform data samples from C classes is denoted as $\mathbf{Y}=[\mathbf{Y}_1,...,\mathbf{Y}_C]\in\Re^{M\times CK}$. Every class $c\in\mathcal{C}=\{1,...,C\}$ has K samples, $\mathbf{Y}_c=[\mathbf{y}_{c,1},...,\mathbf{y}_{c,K}]\in\Re^{M\times K}$, $\mathbf{Y}_{\setminus c}=[\mathbf{Y}_1,...,\mathbf{Y}_{c-1},\mathbf{Y}_{c+1},...,\mathbf{Y}_C]$ and $\mathbf{Y}_{\setminus\{c,k\}}=[\mathbf{Y}_1,...,[\mathbf{y}_{c,1},...,\mathbf{y}_{c,k-1},\mathbf{y}_{c,k+1},....,\mathbf{y}_{c,K}],...,\mathbf{Y}_C]\in\Re^{M\times(CK-1)}$. The k-th representation from class c, the ℓ_p-norm and the Hadamard product are denoted as $\mathbf{y}_{c,k}\in\Re^M$ ($\forall c\in\mathcal{C},\forall k\in\mathcal{K}=\{1,...,K\}$), $\|.\|_p$ and $\odot$, respectively. $I(\mathbf{x};\mathbf{y})$ denotes the mutual information between vectors $\mathbf{x}$ and $\mathbf{y}$.

A. Model and Learning Approach Overview

Our model and learning approach centers around four elements: (i) joint modeling of two *nonlinear transforms* with (ii) *conditional privacy-utility prior*, which is described with nonlinear transforms that have (iii) a *discriminative prior* and (iv) an *ambiguous prior* in a supervised setup.

Nonlinear Model. Two *nonlinear transforms*[1] are jointly modeled, where a conditional privacy-utility is encoded through a notion for discrimination and ambiguization.

– *Joint Nonlinear Transform Model:* Our joint model is defined as:

$$p\big(\mathbf{Y}_{\{c,k\}},\mathbf{z}_{c,k}|\mathbf{x}_{c,k},\mathbf{W}\big)=\int_{\boldsymbol{\theta}}p\big(\mathbf{Y}_{\{c,k\}},\mathbf{z}_{c,k},\boldsymbol{\theta}|\mathbf{x}_{c,k},\mathbf{W}\big)\,\mathrm{d}\boldsymbol{\theta},\quad(1)$$

where $\mathbf{x}_{c,k}\in\Re^N$ represents the input data, $\mathbf{z}_{c,k}\in\Re^M$ represents the public (protected) representation of $\mathbf{x}_{c,k}$, $\mathbf{Y}_{\{c,k\}}=[\mathbf{y}_{d,\{c,k\}},\mathbf{y}_{a,\{c,k\}}]\in\Re^{M\times 2}$ represents the nonlinear transform representation associated to discrimination and ambiguization, $\boldsymbol{\theta}=\{\boldsymbol{\theta}_d,\boldsymbol{\theta}_a\}$ represents the parameters of model in the corresponding discriminative and ambiguous nonlinear transforms, and $\mathbf{W}=[\mathbf{W}_d,\mathbf{W}_a]$, $\mathbf{W}_d\in\Re^{M\times N}$, $\mathbf{W}_a\in\Re^{M\times N}$ represents the linear maps in the corresponding nonlinear transforms.

– *Model Assumptions:* We focus on the probability $p(\mathbf{x}_{c,k},\mathbf{z}_{c,k},\mathbf{Y}_{\{c,k\}},\boldsymbol{\theta}|\mathbf{W})$ that factors as $p(\mathbf{x}_{c,k}|\mathbf{z}_{c,k},\mathbf{Y}_{\{c,k\}},\boldsymbol{\theta},\mathbf{W})\ p(\mathbf{z}_{c,k},\mathbf{Y}_{\{c,k\}},\boldsymbol{\theta}|\mathbf{W})$ and assume that $p(\mathbf{z}_{c,k},\mathbf{Y}_{\{c,k\}},\boldsymbol{\theta}|\mathbf{W})=p(\mathbf{z}_{c,k},\mathbf{Y}_{\{c,k\}},\boldsymbol{\theta})$.

– *Minimum Information Loss Prior:* To allow adequate coherence and conditioning on the transform matrices $\mathbf{W}_d$ and $\mathbf{W}_a$, a prior $p(\mathbf{W})=p(\mathbf{W}_d)p(\mathbf{W}_a)$ is used.

Discriminative-Ambiguous Setup. In general, knowing the probability distribution for the data $\mathbf{x}_{c,k}$ allows us to use a probability distribution for the transform data $\mathbf{z}_{c,k}=\varphi_x(\mathbf{x}_{c,k})$ and the key (side information) $\mathbf{s}$ and design a privacy-preserving system such that: 1) to simultaneously *minimize* privacy leak and to *maximize* data utility; or 2) to maximize (minimize) data utility (privacy leakage) under constraint on the privacy leakage (data utility). Moreover, knowing the probability distributions and using the mutual information we have a possibility to measure *privacy leakage* by: a) *reconstruction leakage* from a sample $I(\mathbf{x}_{c,k};\mathbf{z}_{c,k})$; b) *database structure leakage* from an encoded data, i.e., $I(\mathbf{x}_{1,1},...,\mathbf{x}_{C,K};\mathbf{z}_{1,1},...,\mathbf{z}_{C,K})$; c) *database structure leakage* around a query $\mathbf{q}_c$, i.e, $I(\mathbf{z}_{1,1},...,\mathbf{z}_{C,K};\mathbf{q}_c)$. At the same time, we have a possibility to measure the *data utility* using a conditional mutual information as $I(\mathbf{z}_{c,k};\mathbf{q}_c|\mathbf{s})$. However, the joint probability distribution function of $\mathbf{x}_{c,k}$, $\mathbf{z}_{c,k}=\varphi_x(\mathbf{x}_{c,k})$ and $\mathbf{s}$ are unknown in advance and we only have a limited amount of observed data. Therefore, we do not have a possibility to use the mutual information as a basis in expressing and solving the related problem.

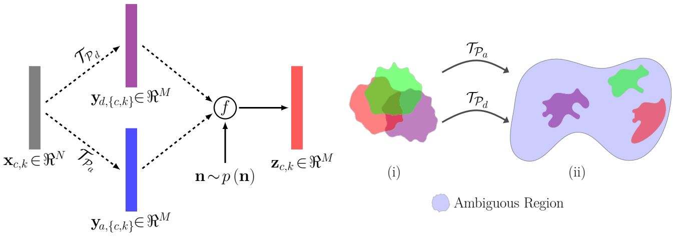

$\mathcal{T}_{\mathcal{P}_d}$
$\mathbf{y}_{d,\{c,k\}}\in\Re^M$
$\mathbf{x}_{c,k}\in\Re^N$
$\mathcal{T}_{\mathcal{P}_a}$
$\mathbf{y}_{a,\{c,k\}}\in\Re^M$
f
$\mathbf{n}\sim p(\mathbf{n})$
$\mathbf{z}_{c,k}\in\Re^M$
$\mathcal{T}_{\mathcal{P}_a}$
$\mathcal{T}_{\mathcal{P}_d}$
(i)
(ii)
Ambiguous Region

Fig. 1: (left panel) Extracting the discriminative representation $\mathbf{y}_{d,\{c,k\}}=\mathcal{T}_{\mathcal{P}_d}(\mathbf{x}_{c,k})$ and ambiguous representation $\mathbf{y}_{a,\{c,k\}}=\mathcal{T}_{\mathcal{P}_a}(\mathbf{x}_{c,k})$ from the corresponding learned nonlinear transforms, and obtaining the final privacy-protected representation $\mathbf{z}_{c,k}$; (right panel) Visualizing three classes of data (i) in the original domain and (ii) in the transform domain.

– *Conditional Privacy-Utility Prior:* Instead of characterizing the privacy-utility trade-off with mutual information, we describe a conditional privacy-utility prior using a *discriminative* nonlinear transform representation $\mathbf{y}_{d,\{c,k\}}$ and an *ambiguous* nonlinear transform representation $\mathbf{y}_{a,\{c,k\}}$. We model the protected template $\mathbf{z}_{c,k}$ and its discrimination and ambigization priors by assuming that the joint probability $p(\mathbf{z}_{c,k},\mathbf{Y}_{\{c,k\}},\boldsymbol{\theta})$ factors as $p(\boldsymbol{\theta}_d|\mathbf{y}_{d,\{c,k\}})\ \ p(\boldsymbol{\theta}_a|\mathbf{y}_{a,\{c,k\}})\ \ p(\mathbf{z}_{c,k}|\mathbf{y}_{d,\{c,k\}},\mathbf{y}_{a,\{c,k\}},\boldsymbol{\theta})$ $p(\mathbf{y}_{d,\{c,k\}})p(\mathbf{y}_{a,\{c,k\}})$. A conditional privacy-utility prior is explicitly modeled as $p(\mathbf{z}_{c,k}|\mathbf{y}_{d,\{c,k\}},\mathbf{y}_{a,\{c,k\}},\boldsymbol{\theta})\propto\exp(-\frac{1}{\beta_Z}\|\mathbf{z}_{c,k},-\mathbf{y}_{d,\{c,k\}}-\mathbf{y}_{a,\{c,k\}}\|_2^2)\exp(-L_{p-u}(\mathbf{y}_{d,\{c,k\}},\mathbf{y}_{a,\{c,k\}}))$, where $L_{p-u}(\mathbf{y}_{d,\{c,k\}},\mathbf{y}_{a,\{c,k\}})$ is the *privacy-utility measure* and β_Z is the scaling parameter.

Note that the proposed modeling allows also to incorporate '*randomness*' in the '*learning stage*' through modeling of the public variable $\mathbf{z}_{c,k}$. This allows us to link our model to the differential privacy machine learning models based on objective perturbation and output perturbation [9], [19].

– *Discrimination Prior:* A *discriminative prior* is modeled as:

$$\begin{aligned}p(\boldsymbol{\theta}_d,\mathbf{y}_{d,\{c,k\}})&=p(\boldsymbol{\theta}_d|\mathbf{y}_{d,\{c,k\}})p(\mathbf{y}_{d,\{c,k\}})\\&\propto\exp(-L_d(\boldsymbol{\theta}_d,\mathbf{y}_{d,\{c,k\}}))p(\mathbf{y}_{d,\{c,k\}}),\quad(2)\end{aligned}$$

where $L_d(\boldsymbol{\theta}_d,\mathbf{y}_{d,\{c,k\}})$ is a discriminative measure for similarity contributions over the parameters $\boldsymbol{\theta}_d=\{\boldsymbol{\theta}_{d1},\boldsymbol{\theta}_{d2}\}=\{\{\boldsymbol{\tau}_{d,1},...,\boldsymbol{\tau}_{d,D1}\},\{\boldsymbol{\nu}_{d,1},...,\boldsymbol{\nu}_{d,D2}\}\}$. The parameters $\boldsymbol{\tau}_{d,d1}$ and

[1] A single nonlinear transform describes a generalized element-wise nonlinearity, that can also be seen as an extension of the sparsifying model [15], [16], [17] and [18].

$\boldsymbol{\nu}_{d,d2}$, $\forall\{d1, d2\} \in \{\{1, ..., D1\}, \{1, ..., D2\}\}$ capture *dissimilar* and *similar* regions in the transform space, respectively. The prior $p(\mathbf{y}_{d,\{c,k\}})$ is a sparsity inducing prior.

– *Ambiguization Prior:* Similarly as in (2), a joint probability models an *ambiguization prior* as:

$$\begin{aligned} p(\boldsymbol{\theta}_a, \mathbf{y}_{a,\{c,k\}}) &= p(\boldsymbol{\theta}_a|\mathbf{y}_{a,\{c,k\}})p(\mathbf{y}_{a,\{c,k\}}) \\ &\propto \exp(-L_a(\boldsymbol{\theta}_a, \mathbf{y}_{a,\{c,k\}}))p(\mathbf{y}_{a,\{c,k\}}), \end{aligned} \quad (3)$$

where $L_a(\boldsymbol{\theta}_a, \mathbf{y}_{a,\{c,k\}})$ is an ambiguization measure for dissimilarity contributions over the parameters $\boldsymbol{\theta}_a=\{\boldsymbol{\theta}_{a1}, \boldsymbol{\theta}_{a2}\}=\{\{\boldsymbol{\tau}_{a,1}, ..., \boldsymbol{\tau}_{a,A1}\}, \{\boldsymbol{\nu}_{a,1}, ..., \boldsymbol{\nu}_{a,A2}\}\}$. Analogously, $\boldsymbol{\tau}_{a,a1}$ and $\boldsymbol{\nu}_{a,a2}, \forall\{a1, a2\} \in \{\{1, ..., A1\}, \{1, ..., A2\}\}$ capture *dissimilar* and *similar* regions in the transform space, respectively.

Learning Strategy. Given training data $\mathbf{X}$, this paper addresses the problem of estimation of the parameters that model the probability $p(\mathbf{Y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}) = p\left(\mathbf{Y}, \mathbf{Z}|\mathbf{X}, \mathbf{W}\right) p\left(\mathbf{W}|\mathbf{X}\right)$, where we assume that $p\left(\mathbf{Y}, \mathbf{Z}|\mathbf{X}, \mathbf{W}\right) = \prod_{k=1}^{K} \prod_{c=1}^{C} p\left(\mathbf{Y}_{\{c,k\}}, \mathbf{z}_{c,k}|\mathbf{x}_{c,k}\mathbf{W}\right)$.

– *Learning Target:* The goal is to estimate the model parameters $\mathbf{z}_{c,k}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}$ and $\mathbf{W}$ which:

(i) maximize the discrimination log-likelihood for the transform representations $\mathbf{y}_{d,\{c,k\}}$,

(ii) maximize the ambiguization log-likelihood for the transform representations $\mathbf{y}_{a,\{c,k\}}$

(iii) under a specific privacy-utility setup.

In other words, given the data, the target is to learn the model parameters that minimize the expected negative log-likelihood of the conditional privacy-utility, *i.e.*, $\mathbb{E}[-\log p(\mathbf{z}_{c,k}|\mathbf{y}_{d,\{c,k\}}, \mathbf{y}_{a,\{c,k\}})]$, but under the priors on $\mathbf{y}_{d,\{c,k\}}$ and $\mathbf{y}_{a,\{c,k\}}$.

– *Empirical Approximation Under Known Labels*: During learning, since we only have access to a finite amount of CK data samples, we approximate the negative likelihood of the conditional privacy-utility using a privacy-utility discrimination and ambiguization measures $L_{p-u}(\mathbf{y}_{d,\{c,k\}}, \mathbf{y}_{a,\{c,k\}})$, $L_d(\boldsymbol{\theta}_d, \mathbf{y}_{d,\{c,k\}})$ and $L_a(\boldsymbol{\theta}_a, \mathbf{y}_{a,\{c,k\}})$.

Moreover, in supervised setup, under known labels, given $\mathbf{Y}_a$ and $\mathbf{Y}_d$ we show an non-parametric equivalent of the expected log-likelihoods $\mathbb{E}[-\log p(\boldsymbol{\theta}_a|\mathbf{y}_{a,\{c,k\}})]$ and $\mathbb{E}[-\log p(\boldsymbol{\theta}_d|\mathbf{y}_{d,\{c,k\}})]$, that is:

$$\begin{aligned} &V(\mathbf{Y}_d, \mathbf{Y}_a) + D(\mathbf{Y}_d) + S(\mathbf{Y}_a) + \delta \sim \\ &\qquad \mathbb{E}\Big[- \log p(\mathbf{z}_{c,k}|\mathbf{y}_{d,\{c,k\}}, \mathbf{y}_{a,\{c,k\}}) \\ &\qquad\qquad p(\boldsymbol{\theta}_d, \mathbf{y}_{d,\{c,k\}})p(\boldsymbol{\theta}_a, \mathbf{y}_{a,\{c,k\}})\Big], \end{aligned}$$

where $V(\mathbf{Y}_d, \mathbf{Y}_a) = \frac{1}{CK}\sum_{c,k}(L_{p-u}(\mathbf{y}_{d,\{c,k\}}, \mathbf{y}_{a,\{c,k\}})$, $D(\mathbf{Y}_d)$ and $S(\mathbf{Y}_a)$ are encoding functions that replace the usage of $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_d$ by the usage of $\mathbf{Y}_a$ and $\mathbf{Y}_d$, respectively, and $\delta \simeq \mathbb{E}[\|\mathbf{z}_{c,k} - \mathbf{y}_{d,\{c,k\}} - \mathbf{y}_{a,\{c,k\}}\|_2^2] + \mathbb{E}[-\log p(\mathbf{y}_{d,\{c,k\}} p(\mathbf{y}_{a,\{c,k\}}))]$.

In general, the '*privacy leakage*' and '*data utility*' measures are quantified according to their definitions in the considered application. In this paper, our goal is not to address an application based quantification. Instead, our objective is to introduce a general model that *jointly learns discriminative and ambiguous representations*, which can be used in the search, identification and recognition applications [4], [6], classification and clustering applications [1], [8], [20], and differentiate-private learning techniques [9], [10], etc. Furthermore, it is worthwhile to mention that obviously both the discriminative and ambiguous representations somehow leak the data, since they are learned from the data. However, we emphasize that we do not impose any randomness in our model that satisfies specific privacy definitions such as differential privacy. This is out of scope for this paper.

– *Learning Algorithm:* We propose an alternating algorithm that efficiently solves an approximative minimization problem over the discriminative and ambiguous representations $\mathbf{y}_{d,\{c,k\}}$ and $\mathbf{y}_{a,\{c,k\}}$, the protected representation $\mathbf{z}_{c,k}$ and the linear maps $\mathbf{W}_d$ and $\mathbf{W}_a$.

II. Joint Nonlinear Transform Modeling

Modeling Multiple Nonlinear Transforms. A joint probability $p(\mathbf{x}_{c,k}, \mathbf{z}_{c,k}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta}, \mathbf{W})$ is considered, where a set of $D_1D_2 + A_1A_2$ nonlinear transform representations associated to discrimination and ambiguization are modeled. At the same time, the nonlinear model is taken into account with discrimination prior $p(\boldsymbol{\theta}_d, \mathbf{y}_{d,\{c,k\}})$, ambiguization prior $p(\boldsymbol{\theta}_a, \mathbf{y}_{a,\{c,k\}})$, sparsity priors $p(\mathbf{y}_{d,\{c,k\}})$ and $p(\mathbf{y}_{a,\{c,k\}})$, and the conditional privacy-utility prior $p(\mathbf{z}_{c,k}|\mathbf{y}_{d,\{c,k\}}, \mathbf{y}_{a,\{c,k\}})$.

Nonlinear Transform Model. The compact description of a single nonlinear transform by a *nonlinear transform model* is defined as:

$$\mathbf{W}_p\mathbf{x}_{c,k} = \mathbf{y}_{p,\{c,k\}} + \mathbf{v}_{p,\{c,k\}}, \quad \mathbf{y}_{p,\{c,k\}} \quad = \mathcal{T}_{\mathcal{P}_p}(\mathbf{x}_{c,k}), \quad (4)$$

where $\mathcal{T}_{\mathcal{P}_p}(\mathbf{x}_{c,k})\colon \Re^N \rightarrow \Re^M$ is the parametric nonlinear function that produces $\mathbf{y}_{p,\{c,k\}}$ by using a set of parameters $\mathcal{P}_p, \forall p \in \{d,a\}$. The term $\mathbf{v}_{p,\{c,k\}}$ is a *nonlinear transform error* vector that represents the deviation of $\mathbf{W}_p\mathbf{x}_{c,k}$ from the targeted transform representation $\mathbf{y}_{p,\{c,k\}} = \mathcal{T}_{\mathcal{P}_p}(\mathbf{x}_{c,k}), \forall p \in \{d, a\}$.

– *Discrimination (Ambiguization) Error Model:* In the simplest form, we model $\mathbf{v}_{p,\{c,k\}}$ as $p(\mathbf{x}_{c,k}|\mathbf{Y}_{\{c,k\}}, \mathbf{W}, \boldsymbol{\theta}) = p(\mathbf{x}_{c,k}|\mathbf{Y}_{\{c,k\}}, \mathbf{W}) \propto \exp(-\Big[\frac{\|\mathbf{W}_d\mathbf{x}_{c,k} - \mathbf{y}_{d,\{c,k\}}\|_2^2}{\beta_{d,0}} + \frac{\|\mathbf{W}_a\mathbf{x}_{c,k} - \mathbf{y}_{a,\{c,k\}}\|_2^2}{\beta_{a,0}}\Big])$, where $\beta_{d,0}$ and $\beta_{a,0}$ are the scaling parameters. Nevertheless, additional knowledge about $\mathbf{v}_{p,\{c,k\}} = \mathbf{W}_p\mathbf{x}_{c,k} - \mathbf{y}_{p,\{c,k\}}$ can be used to model $p(\mathbf{x}_{c,k}|\mathbf{Y}_{\{c,k\}}, \mathbf{W})$.

– *Constrained Information Loss:* Additionally, we have a priors on $\mathbf{W}_d$ and $\mathbf{W}_a$ that penalizes the information loss in order to avoid trivially unwanted matrices, *i.e.*, matrices that have repeated or zero rows. The prior measure is denoted as $\Omega(\mathbf{W}_p) = (\frac{1}{\beta_{p,3}}\|\mathbf{W}_p\|_F^2 + \frac{1}{\beta_{p,4}}\|\mathbf{W}_p\mathbf{W}_p^T - \mathbf{I}\|_F^2 - \frac{1}{\beta_{p,5}}\log|\det \mathbf{W}_p^T\mathbf{W}_p|), \forall p \in \{d, a\}$. This prior is used to regularize the conditioning and the expected coherence of $\mathbf{W}_p$ (for more details please see [21]).

III. Priors Measure Modeling

This section first defines the similarity, dissimilarity and strength measures on the support intersection between two representations. Next, it defines the measures for the conditional privacy-utility, discrimination and ambiguization priors.

Support Intersection Based Measures. Given two representations $\mathbf{z}_{c,k}$ and $\mathbf{z}_{c1,k1}$, note that when $\mathbf{z}_{c,k}^T \mathbf{z}_{c1,k1}$ is considered, $\|\mathbf{z}_{c,k}^+ \odot \mathbf{z}_{c1,k1}^+\|_1 + \|\mathbf{z}_{c,k}^- \odot \mathbf{z}_{c1,k1}^-\|_1$ captures the contribution for the similarity, whereas $\|\mathbf{z}_{c,k}^+ \odot \mathbf{z}_{c1,k1}^-\|_1 + \|\mathbf{z}_{c,k}^- \odot \mathbf{z}_{c1,k1}^+\|_1$ captures the contribution for the dissimilarity between $\mathbf{z}_{c,k}$ and $\mathbf{z}_{c1,k1}$.

The measure Sim associated to a similarity between two representations $\mathbf{y}_{d,\{c,k\}}$ and $\mathbf{y}_{d,\{c1,k1\}}$ is defined as $\mathsf{Sim}(\mathbf{y}_{d,\{c,k\}}, \mathbf{y}_{d,\{c1,k1\}}) = \|\mathbf{y}_{d,\{c,k\}}^- \odot \mathbf{y}_{d,\{c1,k1\}}^-\|_1 + \|\mathbf{y}_{d,\{c,k\}}^+ \odot \mathbf{y}_{d,\{c1,k1\}}^+\|_1$, where $\mathbf{y}_{d,\{c,k\}} = \mathbf{y}_{d,\{c,k\}}^+ - \mathbf{y}_{d,\{c,k\}}^-$,$\mathbf{y}_{d,\{c1,k1\}} = \mathbf{y}_{d,\{c1,k1\}}^+ - \mathbf{y}_{d,\{c1,k1\}}^-$, $\mathbf{y}_{d,\{c,k\}}^+ = \max(\mathbf{y}_{d,\{c,k\}}, \mathbf{0})$ and $\mathbf{y}_{d,\{c,k\}}^- = \max(-\mathbf{y}_{d,\{c,k\}}, \mathbf{0})$.

In a similar way, the measure Dis associated to a dissimilarity between two representations $\mathbf{y}_{a,\{c,k\}}$ and $\mathbf{y}_{a,\{c1,k1\}}$ is defined as $\mathsf{Dis}(\mathbf{y}_{a,\{c,k\}}, \mathbf{y}_{a,\{c1,k1\}}) = \|\mathbf{y}_{a,\{c,k\}}^+ \odot \mathbf{y}_{a,\{c1,k1\}}^-\|_1 + \|\mathbf{y}_{a,\{c,k\}}^- \odot \mathbf{y}_{a,\{c1,k1\}}^+\|_1$.

The measure $\mathsf{Stg}(\mathbf{z}_{c,k}, \mathbf{z}_{c1,k1}) = \|\mathbf{z}_{c,k} \odot \mathbf{z}_{c1,k1}\|_2^2$ captures the strength on the support intersection between two representations.

Privacy-Utility Measure. The conditional privacy-utility prior is modeled as defined in earlier section. We define its privacy-utility measure as: $L_{p-u}(\mathbf{y}_{d,\{c,k\}}, \mathbf{y}_{a,\{c,k\}}) = \frac{1}{\beta_I}\mathsf{Sim}(\mathbf{y}_{d,\{c,k\}}, \mathbf{y}_{a,\{c,k\}}) + \frac{1}{\beta_S}\mathsf{Stg}(\mathbf{y}_{d,\{c,k\}}, \mathbf{y}_{a,\{c,k\}})$, where β_I and β_S are scaling parameters.

Discrimination/Ambiguization Measure. The discrimination prior is modeled as in (2). The prior $p(\mathbf{y}_{p,\{c,k\}}) \propto \exp\left(-\frac{\|\mathbf{y}_{p,\{c,k\}}\|_1}{\beta_{p,1}}\right)$ is sparsity inducing prior, where $\beta_{p,1}$ is a scaling parameter, $\forall p \in \{d, a\}$.

Using Sim and Stg we define the discrimination/ambiguization measure as:

$$L_p(\boldsymbol{\theta}_p, \mathbf{y}_{p,\{c,k\}}) = \frac{1}{\beta_p} \min_{p1,p2 \in \mathcal{D}} \big(\mathsf{Sim}(\mathbf{y}_{p,\{c,k\}}, \boldsymbol{\tau}_{p,p1}) + \mathsf{Stg}(\mathbf{y}_{p,\{c,k\}}, \boldsymbol{\nu}_{p,p2})\big), \quad (5)$$

where $\boldsymbol{\theta}_p$ are parameters that describe the nonlinear transforms related to discrimination (or ambiguization) and β_p is a scaling parameter. The minimization of cost $L_p(\boldsymbol{\theta}_p, \mathbf{y}_{p,\{c,k\}})$ ensures that $\mathbf{y}_{p,\{c,k\}}$ in the transform domain will be located at the point where (i) the similarity contribution w.r.t. $\boldsymbol{\tau}_{p,p1}$ is the smallest measured w.r.t. Sim and (ii) the strength of the support intersection w.r.t. $\boldsymbol{\nu}_{p,p2}$ is the smallest measured w.r.t. Stg.

– Approximation of Discriminative Prior Measure in Expectation: Note that if the number of parameters $\boldsymbol{\tau}_{d,d2}$ and $\boldsymbol{\nu}_{d,d1}$ equals the number of class labels and if the class label is known, *i.e.*, $c = d1 = d2$, then:

$$L_d(\boldsymbol{\theta}_d, \mathbf{y}_{d,\{c,k\}}) \underset{\text{under known label}}{=} L_d(\boldsymbol{\tau}_{d,c}, \boldsymbol{\nu}_{d,c}, \mathbf{y}_{d,\{c,k\}}). \quad (6)$$

Moreover, instead of using the parameters $\boldsymbol{\tau}_{d,c}$ and $\boldsymbol{\nu}_{d,c}$, we use the label information and the transform representations, and approximate (6) in expectation as:

$$\mathbb{E}[L_d(\boldsymbol{\tau}_{d,c}, \boldsymbol{\nu}_{d,c}, \mathbf{y}_{d,\{c,k\}})] \;\sim\; D(\mathbf{Y}_d),$$

where $D(\mathbf{Y}_d) = \frac{1}{CK}\frac{1}{\beta_d}\sum_c \sum_{c1 \neq c} \sum_k \sum_{k1} \big(\mathsf{Sim}(\mathbf{y}_{d,\{c,k\}}, \mathbf{y}_{d,\{c1,k1\}}) + \mathsf{Stg}(\mathbf{y}_{d,\{c,k\}}, \mathbf{y}_{d,\{c1,k1\}})\big)$.

– Approximation of Ambiguization Prior Measure Expectation: Analogously, if the number of parameters $\boldsymbol{\tau}_{a,a1}$ and $\boldsymbol{\nu}_{a,a2}$ equals the number of class labels and if the class label is known, *i.e.*, $c = a1 = a2$, then:

$$L_a(\boldsymbol{\theta}_a, \mathbf{y}_{a,\{c,k\}}) \underset{\text{under known label}}{=} L_a(\boldsymbol{\tau}_{a,c}, \boldsymbol{\nu}_{a,c}, \mathbf{y}_{a,\{c,k\}}), \quad (7)$$

and in expectation approximate (7) as:

$$\mathbb{E}[L_a(\boldsymbol{\tau}_{a,c}, \boldsymbol{\nu}_{a,c}, \mathbf{y}_{a,\{c,k\}})] \;\sim\; S(\mathbf{Y}_a), \quad (8)$$

where $S(\mathbf{Y}_a) = \frac{1}{CK}\frac{1}{\beta_d}\sum_c \sum_k \sum_{k1 \neq k} \big(\mathsf{Sim}(\mathbf{y}_{a,\{c,k\}}, \mathbf{y}_{a,\{c,k1\}}) + \mathsf{Stg}(\mathbf{y}_{a,\{c,k\}}, \mathbf{y}_{a,\{c,k1\}})\big)$.

IV. Learning the Nonlinear Transforms

The first subsection presents the problem formulation and the second subsection gives the solution to the learning problem.

A. Problem Formulation

Given the available training data set $\mathbf{X}$, maximizing $p(\mathbf{Y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}) = p(\mathbf{Y}, \mathbf{Z}|\mathbf{X}, \mathbf{W})\, p(\mathbf{W}|\mathbf{X})$ over $\mathbf{Z}$, $\mathbf{Y}_d$, $\mathbf{Y}_a$, $\boldsymbol{\theta}_d$, $\boldsymbol{\theta}_a$, $\mathbf{W}_d$ and $\mathbf{W}_a$ is difficult. Instead, we take into account the approximations (7) and (8), and address the following problem:

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{Y}_d,\mathbf{Y}_a,\mathbf{W}_d,\mathbf{W}_a} \sum_{p \in \{d,a\}} &\Big(\frac{1}{2}\|\mathbf{W}_p\mathbf{X} - \mathbf{Y}_p\|_F^2 + \lambda_{p,1}\sum_{c,k}\|\mathbf{y}_{p,\{c,k\}}\|_1\Big) \\ &+ \frac{1}{2}\|\mathbf{Z} - \mathbf{Y}_a - \mathbf{Y}_d\|_F^2 + V(\mathbf{Y}_d, \mathbf{Y}_a) \\ &+ D(\mathbf{Y}_d) + S(\mathbf{Y}_a) + \Omega(\mathbf{W}_d) + \Omega(\mathbf{W}_a), \quad (9)\end{aligned}$$

where $\{\lambda_{d,1}, \lambda_{a,1}\}$ are inversely proportional to the scaling parameters $\{\beta_{d,1}, \beta_{a,1}\}$.

Integrated Marginal Maximization. We highlight that for our model, the solution to (9) is not equivalent to the maximum a posterior (MAP) solution[2], which would be difficult to compute, as it involves integration over $\mathbf{x}_{c,k}$, $\mathbf{y}_{p,\{c,k\}}$ and $\boldsymbol{\theta}$. Instead, we perform an integrated marginal minimization that is addressed with (9) and solved by iteratively marginally maximizing $p(\mathbf{Y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}) = p(\mathbf{Y}, \mathbf{Z}|\mathbf{X}, \mathbf{W})\, p(\mathbf{W}|\mathbf{X})$ in $\mathbf{W}_d$, $\mathbf{W}_a$, $\mathbf{y}_{d,\{c,k\}}$, $\mathbf{y}_{a,\{c,k\}}$ and $\mathbf{z}_{c,k}$. This is equivalent to approximatively 1) maximizing the conditional $p(\mathbf{x}_{c,k}|\mathbf{Y}_{\{c,k\}}, \mathbf{W})$ with prior $p(\mathbf{W}_p)$ over $\mathbf{W}_p$, 2) maximizing the conditional $p(\mathbf{x}_{c,k}|\mathbf{Y}_{\{c,k\}}, \mathbf{W})$ with prior $p(\mathbf{z}_{c,k}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta})$ over $\mathbf{y}_{p,\{c,k\}}$ and 3) maximizing the conditional $p(\mathbf{z}_{c,k}|\mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta})$ over $\mathbf{z}_{c,k}$. The algorithm allows us to find a joint local maximum in $\{\mathbf{W}_d, \mathbf{W}_a, \mathbf{Z}, \mathbf{Y}_d, \mathbf{Y}_a, \boldsymbol{\theta}\}$ for $p(\mathbf{Y}, \mathbf{Z}, \mathbf{W}|\mathbf{X})$, such that the privacy-utilaty, the discrimination and ambiguization prior probabilities are maximized.

Targeting Privacy-Utility Trade-Off Point. Consider three sets of parameters as $\{\lambda_I = 1/\beta_I, \lambda_S = 1/\beta_S\}$, $\{\lambda_d = 1/\beta_d, \lambda_{d,1} = 1/\beta_{d,1}\}$ and $\{\lambda_a = 1/\beta_a, \lambda_{a,1} = 1/\beta_{a,1}\}$. If one fixed parameters of two sets and varied the parameters of other one, then, this reflects to the trade-off between discrimination and ambiguization. At the same time, a particular choice of the parameters allows us to target a specific extreme privacy-utility point.

[2] The MAP estimation problem for our model is identical to (9), but has additional terms that are related to the partition functions of $p(\mathbf{x}_{c,k}|\mathbf{z}_{c,k}, \mathbf{Y}_{c,k}, \boldsymbol{\theta}, \mathbf{W})$ and $p(\mathbf{z}_{c,k}, \mathbf{Y}_{\{c,k\}}, \boldsymbol{\theta})$.

B. Learning Algorithm

We propose an iterative, alternating algorithm with five stages: (i) estimating discriminative representation $\mathbf{y}_{d,\{c,k\}}$, (ii) estimating ambiguous representation $\mathbf{y}_{a,\{c,k\}}$, (iii) estimating public variable $\mathbf{z}_{c,k}$, (iv) and (v) updating the linear maps $\mathbf{W}_d$ and $\mathbf{W}_a$ associated to discrimination and ambiguization.

(i) and (ii): *Estimating Discriminative (or Ambiguous) Representation.* Given the available data samples $\mathbf{X}$, the protected representation $\mathbf{Z}$, the discriminative (or ambiguous) representation $\mathbf{Y}_d$ (or $\mathbf{Y}_a$) and the current estimate of the linear map $\mathbf{W}_d$ (or $\mathbf{W}_a$), the discriminative (or ambiguous) representation estimation problem is formulated as:

$$\min_{\mathbf{Y}_p} \tfrac{1}{2}\|\mathbf{W}_p\mathbf{X} - \mathbf{Y}_p\|_F^2 + \tfrac{1}{2}\|\mathbf{Z} - \mathbf{Y}_d - \mathbf{Y}_a\|_F^2 + V(\mathbf{Y}_d, \mathbf{Y}_a)$$
$$+ D(\mathbf{Y}_d) + S(\mathbf{Y}_a) + \lambda_{p,1}\textstyle\sum_{c,k}\|\mathbf{y}_{p,\{c,k\}}\|_1, \forall p \in \{d, a\}. \quad (10)$$

If $p = d$ then the problem is associated to discriminative representation estimation and if $p = a$ then the problem is associated to ambiguous representation estimation.

Denote $\mathbf{U}_d = \mathbf{W}_d\mathbf{X} + \mathbf{Y}_a - \mathbf{Z}$, $\mathbf{U}_a = \mathbf{W}_a\mathbf{X} + \mathbf{Y}_d - \mathbf{Z}$ and $\mathbf{y}_p = \mathbf{y}_{p,\{c,k\}}$, then problem (10) per $\mathbf{y}_p, \forall p \in \{d, a\}$ reduces to $P_{P-U} : \min_{\mathbf{y}_p} \frac{1}{2}\|\mathbf{u}_{p,\{c,k\}} - \mathbf{y}_p\|_2^2 + \mathbf{g}_{p,\{c,k\}}^T|\mathbf{y}_p| + \lambda_{p,1}\mathbf{1}^T|\mathbf{y}_p| + \mathbf{s}_{p,\{c,k\}}^T(\mathbf{y}_p \odot \mathbf{y}_p)$ and has a closed-form solution as:

$$\mathbf{y}|_{\{p1,p2\}} = \text{sign}(\mathbf{u}_{p,\{c,k\}}) \quad (11)$$
$$\odot \max(|\mathbf{u}_{p,\{c,k\}}| - \mathbf{g}_{p,\{c,k\}} - \lambda_{p,1}\mathbf{1}, \mathbf{0}) \oslash (\mathbf{k}_{p,\{c,k\}}),$$

where $\mathbf{k}_{p,\{c,k\}} = \mathbf{1} + 2\mathbf{s}_{p,\{c,k\}}$ (The proof is omitted due to space limit. We refer the reader to [22] for analogous proof.)

(iii): *Estimating the Protected Variables* $\mathbf{z}_{c,k}$. Given the estimated discriminative and ambiguous representations $\mathbf{y}_{d,\{c,k\}}$ and $\mathbf{y}_{a,\{c,k\}}$, the coupled (protected) variable is estimated as $\mathbf{z}_{c,k} = \mathbf{y}_{d,\{c,k\}} + \mathbf{y}_{a,\{c,k\}}$.

Note that we can impose randomization whether during the learning phase (objective perturbation) or to the final representation (output perturbation). Let consider $\mathbf{z}_{c,k} = f\left(\mathbf{y}_{d,\{c,k\}}, \mathbf{y}_{a,\{c,k\}}, \mathbf{n}\right)$ in general, where $\mathbf{n}$ is a random noise with distribution $p(\mathbf{n})$ and f is randomization mechanism. Some basic examples of f, are $f = \mathbf{y}_{d,\{c,k\}} + \mathbf{y}_{a,\{c,k\}} + \mathbf{n}$, $f = \mathbf{y}_{d,\{c,k\}} + \left(\mathbf{y}_{a,\{c,k\}} \odot \mathbf{n}\right)$, $f = \left(\mathbf{y}_{d,\{c,k\}} + \mathbf{y}_{a,\{c,k\}}\right) \odot \mathbf{n}$.

(iv) and (v): *Updating the Linear Maps* $\mathbf{W}_d$ *and* $\mathbf{W}_a$. Given the data samples $\mathbf{X}$ and the corresponding transform representations $\mathbf{Y}_p, \forall p \in \{d, a\}$, then the problem associated to the estimation of the linear map $\mathbf{W}_p$, reduces to $\min_{\mathbf{W}_p} \|\mathbf{W}_p\mathbf{X} - \mathbf{Y}_p\|_2^2 + \frac{\lambda_{p,3}}{2}\|\mathbf{W}_p\|_F^2 + \frac{\lambda_{p,4}}{2}\|\mathbf{W}_p\mathbf{W}_p^T - \mathbf{I}\|_F^2 - \lambda_{p,5}\log|\det\mathbf{W}_p^T\mathbf{W}_p|$, where $\{\lambda_{p,3}, \lambda_{p,4}, \lambda_{p,5}\}$ are inversely proportional to the scaling parameters $\{\beta_{p,3}, \beta_{p,4}, \beta_{p,5}\}$. We use the approximate closed-form solution as proposed in [21].

V. Evaluation of the Proposed Approach

This section evaluates the advantages and the potential of the proposed algorithm.

A. Data, Setup and Performance Measures

a) Data Sets: The used data sets are AR [23], E-YALE-B [24] and COIL [25]. All the images from the respective datasets were downscaled to resolutions 24×24, 20×25 and 21×21, respectively, and are normalized to unit variance.

b) Algorithm Setup: An on-line variant is used for the update of $\mathbf{W}$ w.r.t. a subset of the available training set. The used batch size is equal to $87\%, 85\%$ and 90% of the total amount of the available training data from the respective datasets AR, E-YALE-B, and COIL. The nonlinear transform (NT) dimension is set to $M = 2100$. The algorithm is initialized with $\mathbf{W}$ and $\boldsymbol{\theta}$ having i.i.d. Gaussian (zero mean, unit variance) entries and is terminated after the 100th iteration. The results are obtained as an average of 5 runs.

B. Numerical Experiments

Summary: The experiments consist of three parts:

– *NT Properties:* In the first series of the experiments, we measure the run time t of the proposed algorithm, the conditioning number $\kappa(\mathbf{W}) = \frac{\sigma_{max}}{\sigma_{min}}$ (ratio of the largest to smallest singular value in the SVD of $\mathbf{W}$) and the expected mutual coherence $\mu(\mathbf{W})$ of the shared linear map $\mathbf{W}$ in the learned NTs.

– *k-NN Classification Performance:* In this part, we also split every databases into training and test set and learn privacy-preserved NTs with the proposed algorithm on the training set. We use the learned NTs to assign a discriminative, ambiguous and privacy-preserved representation for the test data and then preform a k-NN [26] search using the corresponding test discriminative, ambiguous and coupled (protected) representation.

Evaluation Results: The results are shown in Tables I, II, and Figure 2.

– *NT Properties*: The learned NTs for all the data sets have relatively low computational time per iteration. All NT have good conditioning numbers and low expected coherence. The results are given in Table I.

– *k-NN Classification Performance*: The results of the k-NN performance on all databases is shown in Table II. As a baseline we use k-NN on the original data and report improvements of 3.1%, 2.4% and 3.3% over the baseline results for the respective databases.

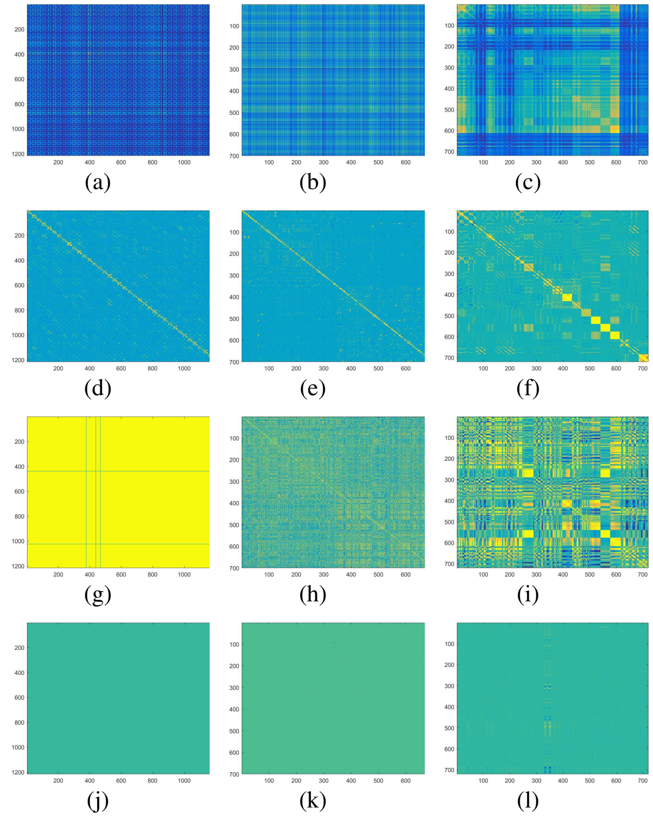
– *Scatter Matrices Comparison*: The scatter matrices $\mathbf{X}^T\mathbf{X}$, $\mathbf{Y}_d^T\mathbf{Y}_d$, $\mathbf{Y}_a^T\mathbf{Y}_a$ and $\mathbf{Y}_d^T\mathbf{Y}$ are depicted in Figure 2, for the considered databases. The results show the capability of proposed model to generate discriminative and ambiguous representations, while there is a high correlation between the representations in the original data domain.

AR					E-YALE-B				
$\mathbf{W}_a$		$\mathbf{W}_d$			$\mathbf{W}_a$		$\mathbf{W}_d$		
κ	μ	κ	μ	t	κ	μ	κ	μ	t
155	0.002	144.98	0.002	8.094	23.32	0.001	11.96	0.001	11.24

TABLE I: The computational efficiency per iteration $t[sec]$ for the proposed algorithm, the conditioning number κ and the expected mutual coherence μ for the linear maps $\mathbf{W}_a$ and $\mathbf{W}_d$.

	COIL	E-YALE-B	AR
Discriminative representation	99.86	94.4	88.57
Ambiguous representation	21.25	2.87	11.14
Coupled representation	96.80	94.81	75.38
Original data	100	81.41	84.57

TABLE II: The k-NN results on the original data and the assigned NT representations.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 2: Scatter matrices associated with E-Yale-B: (a), (d), (g) and (j); AR: (b), (e), (h) and (k); COIL-20: (c), (f), (i) and (l). Figures (a), (b) and (c) depict $\mathbf{X}^T\mathbf{X}$ (original domain). Figures (d), (e) and (f) depic $\mathbf{Y}_d^T\mathbf{Y}_d$ (discriminative representation). Figures (g), (h) and (i) depict $\mathbf{Y}_a^T\mathbf{Y}_a$ (ambiguous representation). Figures (j), (k) and (l) depict $\mathbf{Y}_d^T\mathbf{Y}_a$.

Note that, due to space limit, our goal is just to present the potential of our model in the privacy-protecting techniques. As we discussed before, by imposing randomness in the model, one can measure and compare the privacy and utility in considered application.

VI. Conclusion

In this paper we modeled a privacy-utility cost trade-off that is described by nonlinear transforms related to discrimination and ambiguization. A novel joint model was introduced that includes the nonlinear transforms, the privacy-utility cost, the minimum information loss and discriminative and ambiguous priors. Given an observed data, the model parameters were learned by minimizing an empirical expectation of the model log-likelihood. An efficient solution was proposed using block coordinate descend alternating algorithm.

References

[1] B. I. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft, "Learning in a large function space: Privacy-preserving mechanisms for svm learning," *Journal of Privacy and Confidentiality*, vol. 4, no. 1, p. 4, 2012.

[2] A. Blum, K. Ligett, and A. Roth, "A learning theory approach to noninteractive database privacy," *Journal of the ACM (JACM)*, vol. 60, no. 2, p. 12, 2013.

[3] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," in *Advances in Neural Inf. Processing Sys.*, 2009, pp. 289–296.

[4] B. Razeghi, S. Voloshynovskiy, D. Kostadinov, and O. Taran, "Privacy preserving identification using sparse approximation with ambiguization," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, Rennes, France, December 2017, pp. 1–6.

[5] B. Razeghi and S. Voloshynovskiy, "Privacy-preserving outsourced media search using secure sparse ternary codes," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, Calgary, Canada, April 2018, pp. 1992–1996.

[6] A. Iscen, T. Furon, V. Gripon, M. Rabbat, and H. Jégou, "Memory vectors for similarity search in high-dimensional spaces," *IEEE Transactions on Big Data*, vol. 4, no. 1, pp. 65–77, 2018.

[7] B. Razeghi, S. Voloshynovskiy, S. Ferdowsi, and D. Kostadinov, "Privacy-preserving identification via layered sparse code design: Distributed servers and multiple access authorization," in *26th European Signal Processing Conference (EUSIPCO)*, Italy, September 2018.

[8] M. Barni, P. Failla, R. Lazzeretti, A.-R. Sadeghi, and T. Schneider, "Privacy-preserving ecg classification with branching programs and neural networks," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 2, pp. 452–468, 2011.

[9] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 1069–1109, 2011.

[10] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *IEEE 55th Annual Symp. on Foundations of Com. Sci. (FOCS)*. IEEE, 2014, pp. 464–473.

[11] C. Dwork and J. Lei, "Differential privacy and robust statistics," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*. ACM, 2009, pp. 371–380.

[12] Y. O. Basciftci, Y. Wang, and P. Ishwar, "On privacy-utility tradeoffs for constrained data release mechanisms," in *Information Theory and Applications Workshop (ITA), 2016*. IEEE, 2016, pp. 1–6.

[13] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, 2013.

[14] F. du Pin Calmon, A. Makhdoumi, M. Médard, M. Varia, M. Christiansen, and K. R. Duffy, "Principal inertia components and applications," *IEEE Tran. on Inf. Theory*, vol. 63, no. 8, pp. 5011–5038, 2017.

[15] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. of IEEE*, vol. 98, pp. 1045–1057, 2010.

[16] R. Rubinstein, T. Peleg, and M. Elad, "Analysis k-svd: A dictionary-learning algorithm for the analysis sparse model," *IEEE Transactions on Signal Processing*, vol. 61, no. 3, pp. 661–677, 2013.

[17] R. Rubinstein and M. Elad, "Dictionary learning for analysis-synthesis thresholding," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 5962–5972, 2014.

[18] S. Ravishankar and Y. Bresler, "Doubly sparse transform learning with convergence guarantees," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc.(ICASSP)*. IEEE, 2014, pp. 5262–5266.

[19] A. D. Sarwate and K. Chaudhuri, "Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data," *IEEE signal proc. magazine*, vol. 30, pp. 86–94, 2013.

[20] T. Li, Z. Huang, P. Li, Z. Liu, and C. Jia, "Outsourced privacy-preserving classification service over encrypted data," *Journal of Network and Computer Applications*, 2018.

[21] D. Kostadinov, S. Voloshynovskiy, and S. Ferdowsi, "Learning overcomplete and sparsifying transform with approximate and exact closed form solutions," in *European Workshop on Visual Information Processing (EUVIP)*, Tampere, Finland, November 2018.

[22] D. Kostadinov, B. Razeghi, S.Voloshynovskiy, and S.Ferdowsi, "Learning discrimination specific, self-collaborative and nonlinear model," in *IEEE Internacial Conference on Big Knowlage (ICBK)*, Singapore, November 2018.

[23] A. M. Martinez, "The ar face database," *CVC Technical Report24*, 1998.

[24] 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*, vol. 23, no. 6, pp. 643–660, 2001.

[25] S. A. Nene, S. K. Nayar, H. Murase *et al.*, "Columbia object image library (coil-20)," 1996.

[26] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.