# Orthogonality-Promoting Dictionary Learning via Bayesian Inference

**Lei Luo,**[1] **Jie Xu,**[1,2] **Cheng Deng,**[2] **Heng Huang**[1,3]*
[1]Electrical and Computer Engineering, University of Pittsburgh, USA
[2]School of Electronic Engineering, Xidian University, Xian, Shanxi, China, [3]JDDGlobal.com
lel94@pitt.edu, jie.xu@pitt.edu, chdeng.xd@gmail.com, heng.huang@pitt.edu

## Abstract

Dictionary Learning (DL) plays a crucial role in numerous machine learning tasks. It targets at finding the dictionary over which the training set admits a maximally sparse representation. Most existing DL algorithms are based on solving an optimization problem, where the noise variance and sparsity level should be known as the prior knowledge. However, in practice applications, it is difficult to obtain these knowledge. Thus, non-parametric Bayesian DL has recently received much attention of researchers due to its adaptability and effectiveness. Although many hierarchical priors have been used to promote the sparsity of the representation in non-parametric Bayesian DL, the problem of redundancy for the dictionary is still overlooked, which greatly decreases the performance of sparse coding. To address this problem, this paper presents a novel robust dictionary learning framework via Bayesian inference. In particular, we employ the orthogonality-promoting regularization to mitigate correlations among dictionary atoms. Such a regularization, encouraging the dictionary atoms to be close to being orthogonal, can alleviate overfitting to training data and improve the discrimination of the model. Moreover, we impose Scale mixture of the Vector variate Gaussian (SMVG) distribution on the noise to capture its structure. A Regularized Expectation Maximization Algorithm is developed to estimate the posterior distribution of the representation and dictionary with orthogonality-promoting regularization. Numerical results show that our method can learn the dictionary with an accuracy better than existing methods, especially when the number of training signals is limited.

## Introduction

In the last decades, sparse coding, inspired by the sparsity mechanism of human vision system (Olshausen and Field 1996), has become a significant technique in computer vision and machine learning with many real-world applications such as image classification (Wright et al. 2009), visual tracking (Mei and Ling 2011) and cluster analysis (Elhamifar and Vidal 2009). It models signals as linear combinations of a small number of atoms chosen from a large dictionary by solving an $l_0$ minimization problem. In addition to solid theoretical studies (Candes and Tao 2005), numerous linear models following this line of sparse coding have recently emerged as powerful tools to cope with a variety of estimation tasks, *e.g.*, Collaborative Representation Classifier (CRC) (Zhang, Yang, and Feng 2011), Robust Sparse Coding (RSC) (Yang et al. 2011b), Nuclear norm based Matrix Regression (NMR) (Yang et al. 2017a), capped norm based robust dictionary learning (Jiang, Nie, and Huang 2015), and group sparsity based model (Nie et al. 2010; Yuan, Liu, and Ye 2011).

The dictionary plays an important role in these sparse representation based models. A desired dictionary learned from data often outperforms a set of predefined bases (Guo et al. 2016). As a result, dictionary learning (DL) has received a growing interest and a large number of DL algorithms have been proposed in recent years. K-SVD (Aharon, Elad, and Bruckstein 2006), as a classic DL algorithm, alternates between sparse coding of the examples based on the current dictionary and a process of updating the dictionary atoms to better fit the data. However, K-SVD focuses on only the representational power of the dictionary (or the efficiency of sparse coding under the dictionary) without considering its capability for discrimination. To overcome this limitation, (Zhang and Li 2010) proposed a discriminative K-SVD algorithm to learn an over-complete dictionary from a set of labeled training face images. Different from (Zhang and Li 2010), (Yang et al. 2011a) employed the Fisher discrimination criterion to learn a structured dictionary (FDDL for short). However, FDDL is not able to effectively represent the non-linear changes introduced by the pose variation. Thus (Shekhar et al. 2013) presented a robust supervised method for learning a single dictionary to optimally represent both source and target data.

In practical applications, however, labeling samples is usually expensive and time consuming due to the significant human effort involved. Thus, it is desired to develop semi-supervised or weakly-supervised algorithms for efficiently learning a dictionary. To this end, (Wang et al. 2013) proposed robust semi-supervised dictionary learning model, while (Yang and Chen 2017) explored the discrimination of labeled and unlabeled training data by requiring discriminative representation residual and coefficients. However, these semi-supervised DL methods only modify the objective to include a label fit term that renders the learned dictionary as

*Corresponding Author. L. Luo, H. Huang were partially supported by U.S. NSF IIS 1836945, IIS 1836938, DBI 1836866, IIS 1845666, IIS 1852606, IIS 1838627, IIS 1837956.

discriminative as possible, which may lead to sub-optimal classification performance. To alleviate this shortcoming, (You et al. 2018) considered a weak-supervision setting for analysis dictionary learning that is suitable for classification. priors.

The performance of the methods mentioned above is highly dependent on some prior knowledge such as noise variance and sparsity level (Chen et al. 2013) for choosing a proper regularizer. In practice, nevertheless, these prior information are usually complex and unavailable. To mitigate this limitation, nonparametric Bayesian dictionary learning algorithms (Zhou et al. 2009; 2012) are recently developed. They cast dictionary learning as a factor-analysis problem, with the factor loading corresponding to the dictionary elements (atoms). Then, the model parameters are learned by utilizing nonparametric Bayesian techniques like the beta process (BP) (Paisley and Carin 2009), and the Indian buffet process (IBP) (Ghahramani and Griffiths 2006), which circumvents arduous parameter adjustment task and explains DL models from the statistical perspective. To enhance the discrimination of dictionary, (Akhtar, Shafait, and Mian 2016) adaptively built the association between the dictionary atoms with the class labels such that this association signifies the probability of selection of the dictionary atoms in the expansion of classs-pecific data. Taking the uncertainty of the estimates in the inference process into account, (Serra et al. 2017) presented a novel Bayesian approach for the $l_1$ sparse dictionary learning problem based on K-SVD. To promote the sparsity of the representation, (Yang et al. 2017b) leveraged a Gaussian-inverse Gamma hierarchical prior in modeling.

Although many supervised techniques (Akhtar, Mian, and Porikli 2017) can be integrated into Bayesian DL algorithms, they tried to improve the representation performance of sparse coding by constructing overcomplete dictionaries. This is obviously insufficient since such a strategy often results in high computation cost and ambiguity in corresponding representations. Meanwhile, in many practical applications, we may not learn a satisfactory dictionary due to the limited training samples. Thus, how to eliminate the redundancy among the dictionary atoms to improve the representational power of sparse coding becomes an urgent problem to be solved.

**Our Contributions.** In this paper, we propose a novel Bayesian dictionary learning method. It uses the Orthogonality-Promoting regularization, *i.e.*, BMD regularizer (Xie et al. 2018), to mitigate correlations among the dictionary atoms. This regularizer encourages the dictionary atoms to be close to being orthogonal, which not only can alleviate overfitting to training data, but also improve the discrimination of the model. To facilitate the design of algorithm, we approximate the BMD regularizers with convex functions. Based on the basic framework of DL, we perform a Regularized Expectation Maximization of model parameters with the approximated BMD regularizer on the desired prior distribution. We model dictionaries, representation coefficients and noise under the Hierarchical formulation. Specially, we consider the relationship among elements of each noise vector using the Scale Mixture of the Vector variate Gaussian (SMVG) distribution, which is a long-tail distribution and often is applied to robust modeling. This paper makes three main contributions:

- Based on the Hierarchical formulation, a novel nonparametric Bayesian dictionary learning model is introduced. It uses Orthogonality-Promoting regularization to eliminate the redundancy among the dictionary atoms, leading to the stronger representation for sparse coding.
- To effectively estimate model parameters, a Regularized Expectation Maximization Algorithm is provided, which considers the structural information of the noise.
- Our experiments on four benchmark databases (AR, Extended Yale B, UCF sports action and Caltech-101 databases) show the superior performance of our method in classification tasks.

# Preliminaries and Background

## Notations

The bold capital and bold lowercase symbols are used to represent matrices and vectors, respectively. The transpose of the matrix $\mathbf{M}$ is defined as $\mathbf{M}^T$. tr($\mathbf{M}$) and $|\mathbf{M}|$ (or det($\mathbf{M}$)) denote the trace and determinant of a square matrix $\mathbf{M}$, respectively. $\exp(\cdot)$ denotes the exponential function and $\text{etr}(\cdot) = \exp(\text{tr}(\cdot))$. $\mathcal{R}$ and $\mathcal{R}^{l\times m}$ is the set of all real number and the set of all real $l \times m$-dimensional matrices, respectively. For a matrix $\mathbf{M}$, its $i$-th row is denoted by $\mathbf{m}^i$ and $m_{ij}$ denotes the $(i, j)$-th entry of $\mathbf{M}$. If a $l \times l$ matrix $\mathbf{M}$ is positive semi-define, we denote $\mathbf{M} \succeq 0$ or $\mathbf{M} \in R_l^+$. $\mathcal{E}(\mathbf{M})$ and Cov($\mathbf{M}$) denote the expectation and covariance of $\mathbf{M}$, respectively. $\mathbf{I}_p$ represents a $p \times p$ identity matrix. $\mathbf{z} \sim N_l(0, \Delta_p)$ denotes that the $p$-dimensional vector $\mathbf{z}$ follows Gaussian distribution with zero mean and variance matrix $\Delta_p$. $\parallel \mathbf{z} \parallel_1$ and $\parallel \mathbf{z} \parallel_2$ denote the $l_1$ and $l_2$ of vector $\mathbf{z}$, respectively. $\parallel \mathbf{M} \parallel_F$ defines the Frobenius norm of the matrix $\mathbf{M}$, which is equal to the $l_2$-norm of Vec($\mathbf{M}$), *i.e.*, $\parallel \mathbf{M} \parallel_F = \parallel \text{Vec}(\mathbf{M}) \parallel_2$. $\nabla\phi(\cdot)$ denotes the gradient of function $\phi(\cdot)$.

## Dictionary Learning

Let $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n]$ be a given dictionary, where each atom $\mathbf{a}_i \in \mathcal{R}^d$. Considering the classical sparse coding task, a signal $\mathbf{y} \in \mathcal{R}^d$ can be approximately represented by a linear combination of a few atoms from the dictionary $\mathbf{A}$ as:

$$\mathbf{y} \approx \mathbf{Ax} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n, \qquad (1)$$

where $\mathbf{x} = [x_1, x_2, \cdot\cdot\cdot, x_n] \in \mathcal{R}^n$ is a sparse coefficient vector. To this end, $\mathbf{x}$ is characterized by an $l_0$-norm, which leads to the $l_0$-norm minimization problem.

However, the minimization of $l_0$-norm is an NP hard problem. Donoho proved that "for most large under-determined systems of linear equations, the minimal $l_1$-norm near-solution approximates the sparsest near-solution" (Donoho 2006), therefore recent research usually formulates the sparse coding problem as the minimization of $l_1$-norm, *i.e.,* the coding coefficient can be got by solving the following equation:

$$min_{\mathbf{x}} \parallel \mathbf{y} - \mathbf{Ax} \parallel_2^2 + \alpha \parallel \mathbf{x} \parallel_1 . \qquad (2)$$

| Cases | Squared Frobenius Norm (SFN) | Von Neumann divergence (VND) | Log-Determinant Divergence (LDD) |
|---|---|---|---|
| $\phi(\mathbf{X})$ | $\Vert \mathbf{X} \Vert_F^2$ | $tr(\mathbf{X}\text{log}\mathbf{X}-\mathbf{X})$ | $-\text{logdet}\mathbf{X}$ |
| $\Lambda_\phi(\mathbf{X},\mathbf{Y})$ | $\Vert\mathbf{X}-\mathbf{Y}\Vert_F^2$ | $tr(\mathbf{X}\text{log}\mathbf{X}-\mathbf{X}\text{log}\mathbf{Y}-\mathbf{X}+\mathbf{Y})$ | $tr(\mathbf{X}\mathbf{Y}^{-1})-\text{logdet}(\mathbf{X}\mathbf{Y}^{-1})$ |
| $\Lambda_{\phi,con}(\mathbf{X},\mathbf{Y})$ | $\Vert \mathbf{X}-\mathbf{Y} \Vert_F^2 + tr(\mathbf{X})$ | $tr((\mathbf{X}+\varepsilon\mathbf{Y})\text{log}(\mathbf{X}+\varepsilon\mathbf{Y}))$ | $-\text{logdet}(\mathbf{X}+\varepsilon\mathbf{Y})+(\text{log}\frac{1}{\varepsilon})tr(\mathbf{X})$ |

Table 1: Three different cases for $\phi(\mathbf{X})$, which induce three BMDs, where $\varepsilon > 0$ is a small scalar

where $\alpha > 0$ is a balance parameter. In (2), the first term is called as reconstruction error, and the second term is the sparsity penalty.

The choice of dictionary $\mathbf{A}$ dominates the representation performance of coefficients $\mathbf{x}$. To obtain a better dictionary from the training set of $k$ samples $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdot\cdot\cdot, \mathbf{y}_k] \in R^{d\times k}$, many dictionary learning (DL) algorithms have been proposed, the basic idea of which is to minimize the following empirical cost function over both a dictionary $\mathbf{A}$ and a sparse coefficients matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdot\cdot\cdot, \mathbf{x}_k] \in \mathcal{R}^{d\times k}$:

$$\begin{aligned} &min_{\mathbf{A},\mathbf{X}} \parallel \mathbf{Y} - \mathbf{A}\mathbf{X} \parallel_F^2 + \beta \parallel \mathbf{X} \parallel_1 \\ &s.t., \ \parallel \mathbf{a}_i \parallel_2^2 \leq 1, \ \forall \ i = 1, 2, \cdot\cdot\cdot, n, \end{aligned} \tag{3}$$

where $\beta > 0$ is a balance parameter. The constraint $\parallel \mathbf{a}_i \parallel_2^2 \leq 1$ targets at preventing dictionary $\mathbf{A}$ from being arbitrarily large because it would cause very small values of coefficients matrix $\mathbf{X}$. In (3), one seeks to match the dictionary $\mathbf{A}$ to the imagery of interest, while simultaneously encouraging a sparse representation $\mathbf{X}$.

## Orthogonality-Promoting Regularization

Orthogonality-promoting regularization, preventing the redundancy among the learned variables, has been recently studied in some machine learning problems including latent variable modeling, multitask learning and metric learning (Xie et al. 2018). Due to the easy convex relaxation and complete theoretical guarantee, we choose BMD regularizer as an orthogonality-promoting regularization in this paper.

*Definition 1.* Given a strictly convex, differentiable function $\phi : R^{l\times m} \longrightarrow \mathcal{R}$. For any two real symmetric matrix $\mathbf{X}, \mathbf{Y} \in R^{l\times l}$, a BMD is defined as:

$$\Lambda_\phi(\mathbf{X}, \mathbf{Y}) = \phi(\mathbf{X}) - \phi(\mathbf{Y}) - tr((\nabla\phi(\mathbf{Y}))^T(\mathbf{X} - \mathbf{Y})). \tag{4}$$

Different functions $\phi$ can induce different versions of BMD, which can been used to measure the *closeness* between two matrices. According to (Xie et al. 2018), we summarize three cases about $\phi$ in Table 1. The second line of Table 1 shows three popular $\phi$ functions, which generate three BMDs, *i.e.*, Squared Frobenius Norm (SFN), Von Neumann divergence (VND) and Log-Determinant Divergence (LDD) as displayed in the third line of Table 1. The convex relaxations for three different BMDs are described in the last line of Table 1.

# Proposed Method

In this section, we first describe our Hierarchical model for dictionary learning, then provide a Regularized Stochastic Variational Inference (RSVI) to estimate model parameters.

## Overview of the Proposed Framework

Most of the compressive sensing literature assumes "off-the-shelf" wavelet and DCT bases/dictionaries, but recent denoising and inpainting research has demonstrated the significant advantages of learning an often over-complete dictionary matched to the signals of interest. In previous section, we revisited the basic sparse dictionary learning model. For the convenience, we rewritten it as:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E}, \tag{5}$$

where $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \cdot\cdot\cdot, \mathbf{e}_k]$ is the representation error matrix.

Our goal is to estimate the optimal dictionary $\mathbf{A}$ and representation coefficients matrix $\mathbf{X}$ according to the given prior information. From statistical viewpoint, model (3) assumes each $\mathbf{x}_i$ follows independent identically distributed (i.i.d.) with Laplace distribution, while each error vector $\mathbf{e}_i$ is characterized using independent Gaussian distribution. However, these simple priors cannot cope with those complex data from real-world (Luo et al. 2018). To address this issue, in the following, we develop a hierarchical Bayesian model with orthogonality-promoting regularization for learning dictionaries.

**Modeling error matrix E**. In Bayesian modeling, we need introduce a prior distribution on error matrix $\mathbf{E}$. The scale mixture of the Gaussian distribution (Luo et al. 2018) belongs to the category of the elliptically contoured distribution. Compared with Gaussian distribution, it has heavier tails, which is beneficial for robust modeling. Meanwhile, considering the correlation between elements of the practical noise matrix $\mathbf{E}$, Scale mixture of the Vector variate Gaussian (SMVG) distribution is used to model it in the first layer. That is,

$$\mathbf{e}_i = U_i^{-1/w}\mathbf{\Phi}_i^{-1/2}\mathbf{z}_i, \tag{6}$$

where $\mathbf{z}_i \sim N_d(\mathbf{0}, \mathbf{I}_d)$, $i = 1, 2, \cdots, k$ and $w > 0$. Each $\mathbf{\Phi}_i$ is called precision matrix. Each $\mathbf{e}_i$ is controlled by the nonnegative parameter $U_i$ which is similar to the weight of each group in group sparsity (Yuan, Liu, and Ye 2011).

Setting $w$ as 2, (6) is equivalent to

$$P(\mathbf{e}_i|\mathbf{\Phi}_i, U_i, \mathbf{x}_i) = \frac{|U_i\mathbf{\Phi}_i|}{(2\pi)^{d/2}}exp\left(-\frac{1}{2}\mathbf{e}_i^T(U_i\mathbf{\Phi}_i)\mathbf{e}_i\right). \tag{7}$$

It is worth noting that most of existing methods assume that elements of the practical noise matrix $\mathbf{E}$ are independently generated. But here, we use $U_i\mathbf{\Phi}_i$ to learn the structure of noise vector $\mathbf{e}_i$, which is suitable for image classification with the occlusions or illumination.

Since $\mathbf{E} = \mathbf{Y} - \mathbf{AX}$, (7) is rewritten as

$$\begin{aligned} & P(\mathbf{y}_i|\boldsymbol{\Phi}_i, U_i, \mathbf{x}_i, \mathbf{A}) \\ & = \frac{|U_i\boldsymbol{\Phi}_i|}{(2\pi)^{d/2}} exp\left(-\frac{1}{2}(\mathbf{y}_i - \ \mathbf{Ax}_i)^T(U_i\boldsymbol{\Phi}_i)(\mathbf{y}_i - \ \mathbf{Ax}_i)\right). \end{aligned} \tag{8}$$

Suppose different samples are independent of each other. Then, we have

$$P(\mathbf{Y}|\boldsymbol{\Phi}, \mathbf{U}, \mathbf{X}, \mathbf{A}) = \prod_i^k P(\mathbf{y}_i|\boldsymbol{\Phi}_i, U_i, \mathbf{x}_i, \mathbf{A}). \tag{9}$$

Here $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \cdots, \boldsymbol{\Phi}_k]$ and $\mathbf{U} = [U_1, U_2, \cdots, U_k]$.

In the second layer, we use Jeffrey's prior to fit each scalar variable $U_i$ according to (Luo et al. 2018). Then,

$$P(\mathbf{U}) = \prod_{i=1}^k P(U_i) \propto \prod_{i=1}^k \frac{1}{U_i}. \tag{10}$$

Gamma distribution (Luo et al. 2018) is one of the most widely used prior for the precision matrix $\boldsymbol{\Phi}_i$ of the random effects since it provides a convenient conjugate prior for multivariate normal distribution. Thus, for each super parameter $\boldsymbol{\Phi}_i$, we impose the matrix variate Gamma prior on it, *i.e.*,

$$P(\boldsymbol{\Phi}_i) = \Big(T_d(c)|\mathbf{W}_i|^{-c}\Big)^{-1} |\boldsymbol{\Phi}_i|^{c-\frac{1}{2}(d+1)}\text{etr}(-\mathbf{W}\boldsymbol{\Phi}_i), \tag{11}$$

where $T_d(c)$ is a multivariate gamma function (Luo et al. 2018).

**Modeling dictionary A**. Similarly, to effectively model dictionary $\mathbf{A}$, we use the scale mixture of the Vector variate Gaussian distribution to fit it under Hierarchical formulation, *i.e.*,

$$\mathbf{a}_i = V_i^{-1/w}\boldsymbol{\Psi}_i^{-1/2}\mathbf{g}_i, \tag{12}$$

where $\mathbf{g}_i \sim N_d(\mathbf{0}, \mathbf{I}_d)$, $i = 1, 2, \cdots, k$, and each $\mathbf{a}_i > 0$.

Letting $w = 2$, we have

$$P(\mathbf{a}_i|\boldsymbol{\Psi}_i, V_i) = \frac{|V_i\boldsymbol{\Psi}_i|}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}\mathbf{a}_i^T(V_i\boldsymbol{\Psi}_i)\mathbf{a}_i\right). \tag{13}$$

Assuming that different atoms are independent of each other, we have

$$P(\mathbf{A}|\boldsymbol{\Psi}, \mathbf{V}) = \prod_i^k P(\mathbf{a}_i|\boldsymbol{\Psi}_i, U_i), \tag{14}$$

where $\boldsymbol{\Psi} = [\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \cdots, \boldsymbol{\Psi}_k]$ and $\mathbf{V} = [V_1, V_2, \cdots, V_k]$.

In the second layer, we impose Jeffrey??s prior on each scalar variable $V_i$, *i.e.*,

$$P(\mathbf{V}) = \prod_{i=1}^k P(V_i) \propto \prod_{i=1}^k \frac{1}{V_i}. \tag{15}$$

Meanwhile, matrix variate Gamma prior is chosen as the prior distribution of each $\boldsymbol{\Phi}_i$, *i.e.*,

$$P(\boldsymbol{\Psi}_i) = \Big(T_n(c)|\mathbf{W}_i|^{-c}\Big)^{-1} |\boldsymbol{\Psi}_i|^{c-\frac{1}{2}(n+1)}\text{etr}(-\mathbf{W}\boldsymbol{\Psi}_i). \tag{16}$$

**Modeling coefficients matrix X**. In the first layer, elements in the coefficient vector $\mathbf{x}$ are assumed to be independent and follow the scale mixture of the univariate Gaussian distribution, which has been extensively used to exploit the sparsity of $\mathbf{x}_j$ (Luo et al. 2018). That being said,

$$x_{ij} = (\gamma_{ij})^{-1/2}z_{ij}, \tag{17}$$

where $z_{ij} \sim N(0, 1)$, $(\gamma_{ij})^{-1}$ is the precision of $x_{ij}$, and $x_{ij}$ is the $i$-th element of $\mathbf{x}_j$. This is equivalent to setting

$$P(x_{ij}) = \sqrt{\frac{\gamma_{ij}}{2\pi}} exp(-\gamma_{ij}(x_{ij})^2/2) = N(x_{ij}|0, (\gamma_{ij})^{-1}). \tag{18}$$

Let $\boldsymbol{\gamma}_j = [\gamma_{j1}, \gamma_{j2}, \cdots, \gamma_{jn}]^T$ and $\boldsymbol{\gamma}_j^{\text{diag}} = \text{diag}(\gamma_{1j}, \gamma_{2j}, \cdots, \gamma_{nj})$, then (18) can be re-expressed as:

$$P(\mathbf{x}_j|\boldsymbol{\gamma}_j) = \frac{|\boldsymbol{\gamma}_j|}{(2\pi)^{n/2}} exp\left(-\frac{1}{2}\mathbf{x}_j^T\boldsymbol{\gamma}_j^{\text{diag}}\mathbf{x}_j\right). \tag{19}$$

Thus,

$$P(\mathbf{X}|\boldsymbol{\gamma}) = \Pi_{j=1}^k P(\mathbf{x}_j|\boldsymbol{\gamma}_j), \tag{20}$$

where $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_k]$. The second layer specifies Gamma distributions as hyper priors over each hyper parameters $\boldsymbol{\gamma}_i$ in our method. Therefore,

$$\begin{aligned} P(\boldsymbol{\gamma}_j) &= \Pi_{j=1}^k P(\gamma_{ij}) = \Pi_{j=1}^k \text{Gamma}(\gamma_{ij}|a+1, b) \\ &= \Pi_{j=1}^k \frac{b^{a+1}}{\Gamma(a+1)}\gamma_{ij}^a \text{exp}(-b\gamma_{ij}). \end{aligned} \tag{21}$$

## Regularized Expectation Maximization (REM) Algorithm

The EM algorithm (McLachlan and Krishnan 2007) is a general methodology for maximum likelihood (ML) or MAP estimation. The recent emphasis in the sparse or low-rank reconstruction literature on probabilistic models has led to the increased interest in EM. The EM algorithm starts from an initial guess and iteratively runs an expectation (*E*) step, which evaluates the posterior probabilities using currently estimated parameters, and a maximization (*M*) step, which re-estimates the parameters based on the probabilities calculated in the *E* step. The iterations will not stop until the convergence conditions are satisfied.

In our method, we consider dictionary $\mathbf{A}$ and representation coefficients $\mathbf{X}$ as the hidden variable. Thus, for *E*-step, based on the current parameters, the Maximum-A Posteriori (MAP) estimate of $\mathbf{X}$, denoted as $\hat{\mathbf{X}}$, can be achieved by solving the following problem:

$$\begin{aligned} \hat{\mathbf{X}} &= \text{argmax}_{\mathbf{X}} P(\mathbf{X}|\mathbf{A}, \mathbf{Y}, \boldsymbol{\Phi}, \mathbf{U}, \boldsymbol{\gamma}, \boldsymbol{\Psi}, \mathbf{V}) \\ &= \text{argmax}_{\mathbf{X}} P(\mathbf{Y}|\boldsymbol{\Phi}, \mathbf{U}, \mathbf{X}, \mathbf{A})P(\mathbf{X}|\boldsymbol{\gamma})P(\mathbf{A}|\boldsymbol{\Psi}, \mathbf{V}). \end{aligned} \tag{22}$$

Similarly, the Maximum-A Posteriori (MAP) estimate of $\mathbf{A}$, denoted as $\hat{\mathbf{A}}$, can be achieved by solving the following problem:

$$\begin{aligned} \hat{\mathbf{A}} &= \text{argmax}_{\mathbf{A}} P(\mathbf{A}|\mathbf{X}, \mathbf{Y}, \boldsymbol{\Phi}, \mathbf{U}, \boldsymbol{\gamma}, \boldsymbol{\Psi}, \mathbf{V}) \\ &= \text{argmax}_{\mathbf{A}} P(\mathbf{Y}|\boldsymbol{\Phi}, \mathbf{U}, \mathbf{X}, \mathbf{A})P(\mathbf{X}|\boldsymbol{\gamma})P(\mathbf{A}|\boldsymbol{\Psi}, \mathbf{V}). \end{aligned} \tag{23}$$

Let

$$L_{\mathbf{A}} = P(\mathbf{Y}|\mathbf{\Phi}, \mathbf{U}, \mathbf{X}, \mathbf{A})P(\mathbf{X}|\gamma)P(\mathbf{A}|\mathbf{\Psi}, \mathbf{V}). \tag{24}$$

To encourage the dictionary atoms to be close to being orthogonal, we use BMD regularizers to constrain dictionary atoms. Then, (23) becomes:

$$\hat{\mathbf{A}} = \text{argmin}_{\mathbf{A}} \left( -\text{ln}L_{\mathbf{A}} + \rho\Lambda_{\phi,con}(\mathbf{A}^T\mathbf{A}, \mathbf{I}_n) \right). \tag{25}$$

For Eq. (22), we can obtain a closed-form solution by computing its derivative. However, for Eq. (25), we only can iteratively calculate it. Here we adopt stochastic proximal subgradient method to optimize it, *i.e.*,

$$\begin{aligned}\mathbf{A}^{(t)} \leftarrow &\text{prox}_{\eta^{(t)}R}(\mathbf{A}^{(t-1)} - \eta^{(t)} \\ &\cdot \nabla(\text{ln}L_{\mathbf{A}^{(t-1)}} + \rho\Lambda_{\phi,con}(\mathbf{A}^{(t-1)^T}\mathbf{A}^{(t-1)}, \mathbf{I}_n)),\end{aligned} \tag{26}$$

where

$$\text{prox}_R(\mathbf{B}) = \text{argmin}_{\mathbf{A}\in\mathcal{R}^{d\times n}}\{\frac{1}{2} \parallel \mathbf{A} - \mathbf{B} \parallel_F^2 + R(\mathbf{A})\} \tag{27}$$

is a proximal mapping.

In the *M*-step, using the current posterior probabilities, parameters $\mathbf{\Theta} = \{\gamma, \mathbf{\Phi}, \mathbf{U}, \mathbf{\Psi}, \mathbf{V}\}$ can be obtained by minimizing the following help function:

$$Q(\mathbf{\Theta}, \mathbf{\Theta}^{old}) + \text{log}P(\mathbf{\Theta}), \tag{28}$$

where $\mathbf{\Theta}^{old}$ includes the values of parameters from the previous iteration and $Q(\mathbf{\Theta}, \mathbf{\Theta}^{old})$ can be defined as

$$Q(\mathbf{\Theta}, \mathbf{\Theta}^{old}) = \mathcal{E}_{\mathbf{X}|\mathbf{Y},\mathbf{A},\mathbf{\Theta}^{old}}[\text{log}P(\mathbf{X}, \mathbf{A}, \mathbf{Y}|\mathbf{\Theta})], \tag{29}$$

and

$$P(\mathbf{\Theta}) = P(\gamma)P(\mathbf{\Phi})P(\mathbf{U})P(\mathbf{\Psi})P(\mathbf{V}). \tag{30}$$

Taking the stationary point of the objective function (28) with respect to each parameter, we can obtain their solutions. In fact, we can equivalently write the basic iterative procedure as follows:

$$\mathbf{X} \leftarrow \text{argmax}_{\mathbf{X}}P(\mathbf{X}|\mathbf{A}, \mathbf{Y}, \mathbf{\Phi}, \mathbf{U}, \gamma, \mathbf{\Psi}, \mathbf{V}); \tag{31}$$

$$\hat{\mathbf{A}} \leftarrow \text{argmin}_{\mathbf{A}} \left( -\text{ln}L_{\mathbf{A}} + \rho\Lambda_{\phi,con}(\mathbf{A}^T\mathbf{A}, \mathbf{I}_n) \right); \tag{32}$$

$$\gamma \leftarrow \text{argmax}_{\gamma,\mathbf{\Phi},\mathbf{U},\mathbf{\Psi},\mathbf{V}}P(\gamma, \mathbf{\Phi}, \mathbf{U}, \mathbf{\Psi}, \mathbf{V}). \tag{33}$$

$$\mathbf{\Phi} \leftarrow \text{argmax}_{\gamma,\mathbf{\Phi},\mathbf{U},\mathbf{\Psi},\mathbf{V}}P(\gamma, \mathbf{\Phi}, \mathbf{U}, \mathbf{\Psi}, \mathbf{V}). \tag{34}$$

$$\mathbf{U} \leftarrow \text{argmax}_{\gamma,\mathbf{\Phi},\mathbf{U},\mathbf{\Psi},\mathbf{V}}P(\gamma, \mathbf{\Phi}, \mathbf{U}, \mathbf{\Psi}, \mathbf{V}). \tag{35}$$

$$\mathbf{\Psi} \leftarrow \text{argmax}_{\gamma,\mathbf{\Phi},\mathbf{U},\mathbf{\Psi},\mathbf{V}}P(\gamma, \mathbf{\Phi}, \mathbf{U}, \mathbf{\Psi}, \mathbf{V}). \tag{36}$$

$$\mathbf{V} \leftarrow \text{argmax}_{\gamma,\mathbf{\Phi},\mathbf{U},\mathbf{\Psi},\mathbf{V}}P(\gamma, \mathbf{\Phi}, \mathbf{U}, \mathbf{\Psi}, \mathbf{V}). \tag{37}$$

The prior distribution for each parameter has been given in the previous section. As we know, the complexity of standard EM algorithm for estimating model parameters is very high. To circumvent this shortcoming, we can use stochastic EM (SEM) algorithm (Dombry et al. 2017) to train our model. The basic idea of SEM algorithm is to randomly choose training sample and compute the corresponding solution in each step. But here, we omit the detailed iterative procedure.

## Experiments

We evaluated the performance of the proposed approach on two face data sets: the AR (Martinez 1998) and the Extended YaleB (Lee and J. Ho 2005) for face recognition, a data set for action recognition: UCF sports action (Rodriguez, Ahmed, and Shah 2008) and a data set for object categories: Caltech-101 (Lazebnik, Schmid, and Ponce 2006). These data sets are commonly used in the related literature for evaluation of matrix regression and dictionary learning models. Meanwhile, our method is compared with some representative methods such as Sparse Representation based classification (SRC) (Wright et al. 2009), Collaborative Representation based Classification (CRC) (Zhang, Yang, and Feng 2011), K-SVD (Aharon, Elad, and Bruckstein 2006), Fisher Discrimination Dictionary Learning (FDDL) (Yang et al. 2011a), Label Consistent K-SVD (LS-KSVD) (Jiang, Lin, and Davis 2013), Beta Process Construction (BPC) (Zhou et al. 2009) and Nonparametric Bayesian Correlated Group Regression (BCGR) (Luo et al. 2018). Specifically, SRC and CRC belong to the category of the matrix regression. BCGR is a nonparametric Bayesian matrix regression method. K-SVD, FDDL and LS-KSVD are classical dictionary learning methods. BPC is the well-known Bayesian dictionary learning method, which considers beta process as a prior for learning a dictionary. According to the suggestion of (Luo et al. 2018), we set $a, b, c = 10^{-4}$. To be fair, we adopt the sparse representation based classifier.

### Databases

**The AR face database** contains over 4,000 color face images of 126 people, including frontal views of faces with different facial expressions, lighting conditions and occlusions. The pictures of most persons were taken in two sessions (separated by two weeks). Each section contains 13 color images and 120 individuals (65 men and 55 women) participated in both sessions. The images of these 120 individuals were selected and used in our experiment. We projected $165 \times 120$ cropped face images onto 540-dimensional vectors using a random projection matrix (Wright et al. 2009), thereby extracting Random-Face features.

| Methods | 2/class | 4/class | 6/class | 8/class | 10/class |
|---|---|---|---|---|---|
| SRC | 25.00 | 25.83 | 37.17 | 39.17 | 40.67 |
| CRC | 20.83 | 23.17 | 27.17 | 28.17 | 31.00 |
| K-SVD | 27.83 | 20.83 | 37.17 | 41.83 | 42.50 |
| FDDL | 28.67 | 29.50 | 43.17 | **47.83** | 39.67 |
| LS-KSVD | 25.00 | 21.67 | 36.17 | 44.17 | 43.83 |
| BPC | 16.67 | 19.00 | 26.50 | 32.50 | 26.50 |
| BCGR | 25.91 | 26.82 | 39.03 | 41.21 | 41.67 |
| Our method | **30.95** | **32.13** | **43.78** | 44.29 | **43.86** |

Table 2: Classification accuracy (%) for sunglasses disguise on the AR face Database

**The extended Yale B face database** contains 38 human subjects under nine poses and 64 illumination conditions the

| Methods | 2/class | 4/class | 6/class | 8/class | 10/class |
|---|---|---|---|---|---|
| SRC | 27.17 | 30.00 | 36.83 | 42.50 | 43.67 |
| CRC | 16.33 | 14.33 | 19.00 | 20.17 | 21.83 |
| K-SVD | 23.00 | 16.17 | 30.83 | 36.50 | 41.83 |
| FDDL | 13.50 | 13.50 | 13.67 | 14.17 | 17.50 |
| LS-KSVD | 25.67 | 29.33 | **48.33** | 52.50 | 52.50 |
| BPC | 10.00 | 7.83 | 9.67 | 11.33 | 22.00 |
| BCGR | 27.31 | 30.29 | 35.82 | 43.61 | 44.19 |
| Our method | **31.32** | **34.10** | 47.56 | **52.91** | **53.17** |

Table 3: Classification accuracy (%) for scarf disguise on the AR face Database

light source direction and the camera axis. The 64 images of a subject in a particular pose are acquired at camera frame rate of 30 frames/s, so there is only small change in head pose and facial expression for those 64 images. Here we create 504-dimensional random face features (Wright et al. 2009) from the 192 × 168 cropped face images.

**Caltech-101 database** contains 9, 144 image samples from 101 object categories and a class of background images. The number of samples per class in this database vary between 31 and 800. For classification, we first created 4096- dimensional feature vectors of the images using the 16- layer deep convolutional neural networks for large scale visual recognition. These features were used to create the training and the testing data sets

**UCF sports action database** consists of a set of actions collected from various sports which are typically featured on broadcast television channels such as the BBC and ESPN. The video sequences were obtained from a wide range of stock footage websites including BBC Motion gallery and GettyImages. The dataset includes a total of 150 sequences with the resolution of 720 × 480. The collection represents a natural pool of actions featured in a wide range of scenes and viewpoints. By releasing the data set we hope to encourage further research into this class of action recognition in unconstrained environments. Since its introduction, the dataset has been used for numerous applications such as: action recognition, action localization, and saliency detection. We used the action bank features (Sadanand and Corso 2012) for this database to train and test our approach.

## Experiments on the AR face database

Two groups of experiments are designed on the AR face database. In the first experiment, we test the performance of our method under different number of training samples. As we know, AR face database contains 14 face images without real disguise for each person. We randomly choose 2, 4, 6, 8 or 10 face images from them as training samples. Then, three face images with sunglasses and three face images with scarf from session 1 are considered as test samples, respectively. Tables 2 and 3 report the classification accuracies of SRC, CRC, K-SVD, FDDL, LS-KSVD, BPC, BCGR and our method for the two cases under different number of training samples. The advantage of our method is more obvious with the decreasing of number of training samples. Especially, when there are less than 6 training samples, at least 2.28% improvement is achieved by our methods compared to other method. Although both SRC and BCGR exclusively handle occlusion problem, they rely on overcomplete dictionary. Therefore, for small samples, our method perform better than these two methods.

We conduct the second experiment on session 1 from the AR face database. A random subset with three per subject is chosen to form the training set and the rest are taken as the testing set. The experiment is repeated over five random splits of the data set. The second line of Table 4 lists the average classification accuracies and standard deviations of all methods. It can be observed that the proposed method has a leading performance. It achieves an improvement of more than 2% as compared to the second best method: FDDL. The Bayesian dictionary learning method BPC seems to perform poorly at handling occlusions.

## Experiments on the Extended Yale B database

Similar to the previous experiment, the first {5, 10, 15, 20} face images for each class from the Extended YaleB database are chosen as training set, the rest are taken as the testing set. The experimental results of each method are displayed in Table 5. It can be found that our method gives better classification accuracy. With the decreasing of number of training samples, the advantage of our method is more clear as we expected. Meanwhile, FDDL shows a competitive performance for “20/class” (68.20%). In the second experiment, the Extended YaleB data is randomly split into two subsets with thirty-four persons each. One for training and the other for testing. We repeat the process five times to obtain the results over the five testing sets. The mean classification accuracy and standard deviations over the 5 training set for each algorithm are detailed in the third line of Table 4. It is easy to see that our method is superior to other methods. Meanwhile, SRC (97.85%) and BCGR (97.12%) outperform K-SVD (71.07%) and BPC (83.41%). The classification accuracy of LS-KSVD is 97.53% which is higher than FDDL by 2.47%. On the Extended Yale B database, SRC, LS-KSVD and BCGR is robust to illumination. Their classification accuracies are: 97.85%, 71.07% and 97.12%.

## Experiments on the UCF sports action database

Following a common evaluation protocol (Jiang, Lin, and Davis 2013), we evaluate all methods via five-fold cross validation on the UCF sports action database. The detailed results of SRC, CRC, K-SVD, FDDL, LS-KSVD, BPC, BCGR and our method are shown in the fourth line of Table 4. It can be observed that the proposed method achieves the highest classification result (*i.e.*, 91.38% accuracy), 1.52% improvement over the second best method, SRC. Interestingly, regression based methods such as SRC, CRC and BCGR perform better than dictionary learning methods including K-SVD, LS-KSVD, FDDL, and BPC. The performance of LS-KSVD is not stable in this database. It only achieves a classification accuracy of 71.07%, which is more than ten percentage points lower than the other methods.

| Databases | SRC | CRC | K-SVD | LS-KSVD | FDDL | BPC | BCGR | Our method |
|---|---|---|---|---|---|---|---|---|
| AR | 90.27±1.28 | 72.83±1.67 | 79.23±1.77 | 87.40±1.26 | 89.20±1.37 | 35.52±2.35 | 86.78±1.53 | **91.72**±1.31 |
| Extended Yale B | 97.85±0.41 | 87.15±1.13 | 71.07±1.80 | 97.53±0.71 | 94.79±0.39 | 83.41±0.28 | 97.12±1.06 | **97.87**±1.51 |
| UCF sports | 89.86±2.29 | 87.54±4.87 | 81.16±2.05 | 73.62±8.03 | 83.77±3.30 | 84.35±2.97 | 86.14±2.79 | **91.38**±3.02 |
| Caltech-101 | 97.35±0.17 | 88.28±0.33 | 80.64±2.81 | 97.83±0.23 | 91.35±0.10 | 89.26±0.60 | 92.18±0.56 | **97.91**±0.82 |

Table 4: Classification accuracy (%) and standard deviations on the AR, Extended Yale B database, UCF sports action databases and Caltech-101 database.

| Methods | 5/class | 10/class | 15/class | 20/class |
|---|---|---|---|---|
| SRC | 52.83 | 64.60 | 65.4 | 67.96 |
| CRC | 50.13 | 59.05 | 59.00 | 57.80 |
| K-SVD | 48.65 | 46.12 | 41.54 | 49.4 |
| FDDL | 48.52 | 61.85 | 65.56 | **68.20** |
| LS-KSVD | 53.51 | 59.14 | 60.68 | 64.87 |
| BPC | 43.57 | 55.26 | 26.50 | 40.69 |
| BCGR | 53.29 | 65.03 | 65.28 | 67.23 |
| Our method | **55.42** | **67.31** | **66.07** | 67.13 |

Table 5: Classification accuracy (%) on the Extended Yale Database

### Experiments on the Caltech-101 database

In this section, we implement an experiment on the Caltech-101 database for object recognition. For this data set, we directly use the 3000-dimensional Spatial Pyramid Features of the images provided by Jiang *et al.* (Jiang, Lin, and Davis 2013). From these features, 15 random samples per class are used for training and the remaining samples for testing. The experimental results on this database are summarized in the last line of Table 4. It is seen that the performance of our method is similar to that of SRC and LS-KSVD. Especially, LS-KSVD performs better than K-SVD, which indicates the label information may be beneficial for classification task. BPC (84.35%), as a Bayesian dictionary learning method, has the higher accuracy as compared to K-SVD.

## Conclusions

This paper proposed a non-parametric Bayesian approach for learning a desired dictionary. We hierarchically model each parameter and estimate them using a Regularized Expectation Maximization Algorithm. To eliminate the redundancy of the dictionary atoms, we force a orthogonality-promoting regularization on the dictionary matrix, which improves the performance of sparse coding and the discrimination of the model. A series of experiments on four benchmark databases demonstrate that the proposed model can effectively cope with the practical classification problem, particularly for the case where there is a limited number of training samples.

## References

Aharon, M.; Elad, M.; and Bruckstein, A. 2006. $rmk$-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing* 54(11):4311–4322.

Akhtar, N.; Mian, A. S.; and Porikli, F. 2017. Joint discriminative bayesian dictionary and classifier learning. In *CVPR*, 3919–3928.

Akhtar, N.; Shafait, F.; and Mian, A. 2016. Discriminative bayesian dictionary learning for classification. *IEEE transactions on pattern analysis and machine intelligence* 38(12):2374–2388.

Candes, E. J., and Tao, T. 2005. Decoding by linear programming. *IEEE transactions on information theory* 51(12):4203–4215.

Chen, B.; Polatkan, G.; Sapiro, G.; Blei, D.; Dunson, D.; and Carin, L. 2013. Deep learning with hierarchical convolutional factor analysis. *IEEE transactions on pattern analysis and machine intelligence* 35(8):1887–1901.

Dombry, C.; Genton, M. G.; Huser, R.; and Ribatet, M. 2017. Full likelihood inference for max-stable data. *arXiv preprint arXiv:1703.08665*.

Donoho, D. L. 2006. For most large underdetermined systems of linear equations the minimal ??1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59(6):797–829.

Elhamifar, E., and Vidal, R. 2009. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2790–2797. IEEE.

Ghahramani, Z., and Griffiths, T. L. 2006. Infinite latent feature models and the indian buffet process. In *Advances in neural information processing systems*, 475–482.

Guo, J.; Guo, Y.; Kong, X.; Zhang, M.; and He, R. 2016. Discriminative analysis dictionary learning. In *AAAI*, 1617–1623.

Jiang, Z.; Lin, Z.; and Davis, L. S. 2013. Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE transactions on pattern analysis and machine intelligence* 35(11):2651–2664.

Jiang, W.; Nie, F.; and Huang, H. 2015. Robust dictionary learning with capped l1-norm. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *null*, 2169–2178. IEEE.

Lee, K., and J. Ho, D. K. 2005. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. on PAMI* 27(5):684–698.

Luo, L.; Yang, J.; Zhang, B.; Jiang, J.; and Huang, H. 2018. Nonparametric bayesian correlated group regression with applications to image classification. *IEEE Transactions on Neural Networks and Learning Systems* (99):1–15.

Martinez, A. M. 1998. The ar face database. *CVC Technical Report24*.

McLachlan, G., and Krishnan, T. 2007. *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Mei, X., and Ling, H. 2011. Robust visual tracking and vehicle classification via sparse representation. *IEEE transactions on pattern analysis and machine intelligence* 33(11):2259–2272.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint ?2, 1-norms minimization. In *Advances in neural information processing systems*, 1813–1821.

Olshausen, B. A., and Field, D. J. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):607.

Paisley, J., and Carin, L. 2009. Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 777–784. ACM.

Rodriguez, M. D.; Ahmed, J.; and Shah, M. 2008. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

Sadanand, S., and Corso, J. J. 2012. Action bank: A high-level representation of activity in video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 1234–1241. IEEE.

Serra, J. G.; Testa, M.; Molina, R.; and Katsaggelos, A. K. 2017. Bayesian k-svd using fast variational inference. *IEEE Transactions on Image Processing* 26(7):3344–3359.

Shekhar, S.; Patel, V. M.; Nguyen, H. V.; and Chellappa, R. 2013. Generalized domain-adaptive dictionaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 361–368.

Wang, H.; Nie, F.; Cai, W.; and Huang, H. 2013. Semi-supervised robust dictionary learning via efficient l-norms minimization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1145–1152.

Wright, J.; Yang, A. Y.; Ganesh, A.; Sastry, S. S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence* 31(2):210–227.

Xie, P.; Wu, W.; Zhu, Y.; and Xing, E. P. 2018. Orthogonality-promoting distance metric learning: Convex relaxation and theoretical analysis. *arXiv preprint arXiv:1802.06014*.

Yang, M., and Chen, L. 2017. Discriminative semi-supervised dictionary learning with entropy regularization for pattern classification. In *AAAI*, 1626–1632.

Yang, M.; Zhang, L.; Feng, X.; and Zhang, D. 2011a. Fisher discrimination dictionary learning for sparse representation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 543–550. IEEE.

Yang, M.; Zhang, L.; Yang, J.; and Zhang, D. 2011b. Robust sparse coding for face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 625–632. IEEE.

Yang, J.; Luo, L.; Qian, J.; Tai, Y.; Zhang, F.; and Xu, Y. 2017a. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE transactions on pattern analysis and machine intelligence* 39(1):156–171.

Yang, L.; Fang, J.; Cheng, H.; and Li, H. 2017b. Sparse bayesian dictionary learning with a gaussian hierarchical model. *Signal Processing* 130:93–104.

You, Z.; Raich, R.; Fern, X. Z.; and Kim, J. 2018. Weakly supervised dictionary learning. *IEEE Transactions on Signal Processing* 66(10):2527–2541.

Yuan, L.; Liu, J.; and Ye, J. 2011. Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems*, 352–360.

Zhang, Q., and Li, B. 2010. Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2691–2698. IEEE.

Zhang, L.; Yang, M.; and Feng, X. 2011. Sparse representation or collaborative representation: Which helps face recognition? In *Computer vision (ICCV), 2011 IEEE international conference on*, 471–478. IEEE.

Zhou, M.; Chen, H.; Ren, L.; Sapiro, G.; Carin, L.; and Paisley, J. W. 2009. Non-parametric bayesian dictionary learning for sparse image representations. In *Advances in neural information processing systems*, 2295–2303.

Zhou, M.; Chen, H.; Paisley, J.; Ren, L.; Li, L.; Xing, Z.; Dunson, D.; Sapiro, G.; and Carin, L. 2012. Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images. *IEEE Transactions on Image Processing* 21(1):130–144.