This CVPR2014 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Latent Dictionary Learning for Sparse Representation based Classification

Meng Yang, Shenzhen University, ETH Zurich
Dengxin Dai, ETH Zurich
Linlin Shen, Shenzhen University
Luc Van Gool, ETH Zurich, K.U. Leuven
{yang,dai,vangool}@vision.ee.ethz.ch, llshen@szu.edu.cn

Abstract

Dictionary learning (DL) for sparse coding has shown promising results in classification tasks, while how to adaptively build the relationship between dictionary atoms and class labels is still an important open question. The existing dictionary learning approaches simply fix a dictionary atom to be either class-specific or shared by all classes beforehand, ignoring that the relationship needs to be updated during DL. To address this issue, in this paper we propose a novel latent dictionary learning (LDL) method to learn a discriminative dictionary and build its relationship to class labels adaptively. Each dictionary atom is jointly learned with a latent vector, which associates this atom to the representation of different classes. More specifically, we introduce a latent representation model, in which discrimination of the learned dictionary is exploited via minimizing the within-class scatter of coding coefficients and the latent-value weighted dictionary coherence. The optimal solution is efficiently obtained by the proposed solving algorithm. Correspondingly, a latent sparse representation based classifier is also presented. Experimental results demonstrate that our algorithm outperforms many recently proposed sparse representation and dictionary learning approaches for action, gender and face recognition.

1. Introduction

With the inspiration of sparse coding mechanism of human vision system [3][4], sparse coding by representing a signal as a sparse linear combination of representation bases (i.e., a dictionary of atoms) has been successfully applied to image restoration [1][2], image classification [5][6], to name a few. The dictionary, which should faithfully and discriminatively represent the encoded signal, plays an important role in the success of sparse representation [28]. Taking off-the-shelf bases (e.g., wavelets) as the dictionary [7] might be universal to all types of images but will not be effective enough for specific tasks (e.g., face classification). Instead, learning the desired dictionary from the training data by the latest advances in sparse representation has led to state-of-the-art results in many practical applications, such as image reconstruction [1] [8] [9], face recognition [10][11] [12][21][36], and image classification [8][13][14][15].

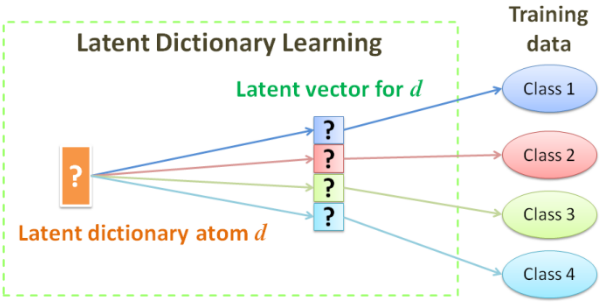

Figure 1: In latent dictionary learning, each dictionary atom $\boldsymbol{d}$ and its associated latent vector are jointly learned, where the latent vector indicates the relatioship between $\boldsymbol{d}$ and class labels.

Current prevailing dictionary learning (DL) approaches can be divided into two main categories: unsupervised dictionary learning and supervised dictionary learning. One representative unsupervised DL approach is the KSVD algorithm [16], which learns an over-complete dictionary of atoms from a set of unlabeled natural image patches. Unsupervised DL methods have been widely applied to image processing tasks, such as image denoising [1][8][9][16], super-resolution [2], and image compression [17]. Besides, in the feature coding of image representation, unsupervised learned dictionary or codebook of local appearance descriptor (e.g., SIFT) has also achieved state-of-the-art performance [6][18].

Without using the label information of training data, the unsupervised dictionary learning method can only require training samples to be sparsely represented by the learned dictionary. The unsupervised learned dictionary is powerful for data reconstruction, but not advantageous for classification tasks. With the class labels of training samples available, the supervised DL methods could exploit the class discrimination information in learning dictionary and thus the learned dictionary has resulted in better classification performance [8][11] [19][20][36].

In the supervised dictionary learning, the discrimination could be exploited from the coding coefficients, the dictionary, or both. Instead of using the standard l_0/l_1-norm sparsity, group sparisity [24] was adopted to regularize the coding coefficients to make the sparse coding coefficients within the same class similar. The discrimination of coding coefficients has also been exploited by learning a dictionary and a classifier over the coding coefficients jointly [8][10][11][19]. Due to the promising performance of class-specific dictionary representation reported in [5], regularizations associated to the dictionary, e.g., reducing the dictionary coherence [20], requiring a sub-dictionary representing well for some class but bad for all the other classes [22][26], has also been introduced in the dictionary updating. In addition, stronger discrimination has been exploited in Fisher discrimination dictionary learning where both discriminative reconstruction error and sparse coefficients were achieved [21][36].

Although improved performance has been reported in the existing dictionary learning approaches, there still remains one critical issue, i.e., how to adaptively build the relationship between dictionary atoms and class labels. In the existing supervised dictionary learning approaches, the label of dictionary atom is predefined and fixed — each dictionary atom is either associated to all classes in [8][10][11][19][24], or assigned to a single class in [20][21][22][26][36]. It is popular to set the label of dictionary atom beforehand; however, this predefined relationship may not be accurate due to the fact that atoms are being updated. In addition, in the case that each dictionary atom has only a single class label, the possible big correlation between different-class dictionary atoms would reduce the discrimination of the learned dictionary. In the case that each dictionary atom is shared by all classes, the mixed information from different classes may reduce the discrimination of the learned dictionary.

In this paper we propose a new discriminative latent dictionary learning (LDL) model to learn a dictionary and a latent matrix jointly, where the latent matrix indicates the relationship between dictionary atoms and class labels, as shown in Fig. 1. The latent matrix is adaptively updated so that it more flexibly associates dictionary atoms to their classes. Meanwhile, the within-class similarity of coding coefficients is enhanced, and a latent-value weighted dictionary coherence term is proposed to reduce the correlation of dictionary atoms between different classes. To this end, the latent correspondence of dictionary atoms to class labels is adaptively built and a latent dictionary is discriminatively learned. Correspondingly, a latent classification model was presented to fully exploit the discrimination of the learned latent dictionary. The LDL is evaluated on the application of action, gender, and face classification. Compared with other state-of-the-art dictionary learning methods, LDL has better or competitive performance in various classification tasks.

The rest of this paper is organized as follows. Section 2 briefly introduces related work. Section 3 presents the proposed LDL model. Section 4 describes the optimization procedure of LDL. Section 5 conducts experiments, and Section 6 concludes the paper.

2. Related Work

Based on predefined relationship between dictionary atoms and class labels, current supervised dictionary learning can be categorized into three main types: shared dictionary learning (i.e., each dictionary atom is associated to all classes), class-specific dictionary learning (i.e., each dictionary atom is assigned to only a single class), and hybrid dictionary (i.e., combination of shared dictionary atoms and class-specific dictionary atoms) learning.

In the first category, a dictionary shared by all classes is learned while the discrimination of coding coefficients is exploited [8][10][11][19][24]. It is very popular to learn a shared dictionary and a classifier over the representation coefficients jointly. Marial *et al.* [19] proposed to learn discriminative dictionaries with a linear classifier of coding coefficients simultaneously. Based on KSVD [16], Zhang and Li [10] also proposed a joint learning algorithm called discriminative KSVD (DKSVD) for face recognition, followed by the work proposed by Jiang *et al*. [11] via adding a label consistent term. Recently, Mairal *et al.* [8] proposed a task-driven DL framework which minimizes different risk functions of the representation coefficients for different tasks. Generally speaking, by fixing all dictionary atoms associated to all classes, a shared dictionary and a classifier over the coding coefficients are jointly learned in the work above [8][10][11][19]. Although a shared dictionary usually has a small size, making the coding in the testing phase efficiently, no class-specific representation residuals can be used.

In the class-specific dictionary learning, each dictionary atom is predefined to correspond to a single class label so that the class-specific reconstruction error could be used for classification [20][21][22][26][36]. Mairal *et al.* [22] introduced a discriminative reconstruction penalty term in the KSVD model [16] for the application of texture segmentation and scene analysis. Yang *et al.* [21] introduced Fisher discrimination both in the sparse coding coefficients and class-specific representation. Castrodad and Sapiro [26] learned a set of action-specific dictionaries with non-negative penalty on both dictionary atoms and representation coefficients. To encourage the dictionaries representing different classes to be as independent as possible, Ramirez *et al.* [20] introduced an incoherence promoting term to the DL model. Since each dictionary atom is fixed to a single class label, the representation residual associated with each class-specific dictionary could be used to do classification; however, the dictionary atoms belonging to different classes can still have big correlations, and the size of the learned dictionary can be very big when there are a large number of classes.

Very recently, the hybrid dictionary combing class-specific dictionary atoms and shared dictionary atoms have been learned. Zhou *et al.* [13] learned a hybrid dictionary with a Fisher-like regularization on the coding coefficients, while Kong *et al.* [12] learned a hybrid dictionary by introducing a coherence penalty term of different sub-dictionaries. Instead of using a flat category structure, Shen *et al.* [27] proposed to learn a dictionary with a hierarchical category structure. Although the shared dictionary atoms could reduce the size of the learned hybrid dictionary to some extent, the shared part and class-specific part also need to be predefined and how to balance these two parts in the hybrid dictionary is not a trivial task.

3. Latent Dictionary Learning (LDL)

Instead of predefining the labels of each learned dictionary atom like the shared dictionary atom and class-specific dictionary atom, we propose a latent dictionary learning model, where the relationship between a dictionary atom and a class label is indicated by a latent value. The learned latent dictionary includes a dictionary $\boldsymbol{D}$=[$\boldsymbol{d}_1$, $\boldsymbol{d}_2$, …, $\boldsymbol{d}_N$] and a latent matrix $\boldsymbol{W}$=[$\boldsymbol{w}_1$,$\boldsymbol{w}_2$,…,$\boldsymbol{w}_C$], where N is the number of dictionary atoms, C is the number of classes, $\boldsymbol{d}_m$ is a dictionary atom, and $\boldsymbol{w}_j$=[$w_{j,1}$,$w_{j,2}$,…,$w_{j,N}$]$^T\in\Re^{N\times 1}$ is a latent vector to indicate the relationship of all dictionary atoms to the j-th class data. For instance, $w_{j,m}$=0 indicates that $\boldsymbol{d}_m$ does not represent the j-th class data; and $w_{j,m}$>0 indicates that $\boldsymbol{d}_m$ is a representation basis of j-th class data.

3.1. Latent dictionary learning model

Denote by $\boldsymbol{A}$=[$\boldsymbol{A}_1$, $\boldsymbol{A}_2$, …, $\boldsymbol{A}_C$] the set of training samples, where $\boldsymbol{A}_j$ is the sub-set of the training samples from the j-th class. Denote by $\boldsymbol{X}$ = [$\boldsymbol{X}_1$, $\boldsymbol{X}_2$, …, $\boldsymbol{X}_C$], where $\boldsymbol{X}_j$ is the sub-matrix containing the coding coefficients of $\boldsymbol{A}_j$ over $\boldsymbol{D}$. Different from the existing sparse representation, we introduce a latent representation model to code $\boldsymbol{A}$ on the desired dictionary $\boldsymbol{D}$. Take the latent representation of j-th class data, $\boldsymbol{A}_j$, as an example. With the latent vector $\boldsymbol{w}_j$ indicating the relationship between $\boldsymbol{D}$ and j-th class data, the latent representation model requires $\boldsymbol{D}$ should well represent all training samples of j-th class, i.e., $\boldsymbol{A}_j\approx\boldsymbol{D}\mathrm{diag}(\boldsymbol{w}_j)\boldsymbol{X}_j$, where diag($\boldsymbol{w}_j$) is a diagonal matrix with $\boldsymbol{w}_j$ as its diagonal vector. In order to make the latent value have physical meaning, the latent value is required to be non-negative. To balance the latent data representation of different classes, the summarization of latent values for each class should be equal to each other, i.e., $\sum_m w_{j,m}=\sum_m w_{l,m}$ for $j\neq l$. Besides, $\boldsymbol{D}$ is also required to have powerful classification ability for $\boldsymbol{A}$. To this end, we propose the following latent dictionary learning (LDL) model:

$$\begin{aligned}\min_{\boldsymbol{D},\boldsymbol{X},\boldsymbol{W}} & \sum_{j=1}^{C}\left\|\boldsymbol{A}_j-\boldsymbol{D}\mathrm{diag}\left(\boldsymbol{w}_j\right)\boldsymbol{X}_j\right\|_F^2+\lambda_1\left\|\boldsymbol{X}_j\right\|_1+\lambda_2\left\|\boldsymbol{X}_j-\boldsymbol{M}_j\right\|_F^2\\ & +\lambda_3\sum\nolimits_{j=1}^{C}\sum\nolimits_{l\neq j}\sum\nolimits_{n=1}^{N}\sum\nolimits_{m\neq n}w_{j,m}\left(\boldsymbol{d}_m^T\boldsymbol{d}_n\right)^2 w_{l,n}\\ & \text{s.t. } w_{j,m}\geq 0\ \ \forall\, j,m;\qquad \sum\nolimits_m w_{j,m}=\delta,\ \ \forall\, m;\end{aligned} \tag{1}$$

where δ is a temporary scalar, which is determined by the initialization of $\boldsymbol{W}$. The initialization of $\boldsymbol{W}$ need ensure the balanced representation of different classes.

In the latent dictionary learning model, the discrimination is exploited via the dictionary itself and the coding coefficients associated to $\boldsymbol{D}$. Apart from requiring the coding coefficient should be sparse, we also minimize the within-class scatter of coding coefficients, $\|\boldsymbol{X}_j$-$\boldsymbol{M}_j\|$, to make the training samples from the same class have similar coding coefficients, where $\boldsymbol{M}_j$ is the mean vector matrix with the same size as $\boldsymbol{X}_j$ and takes the mean column vector of $\boldsymbol{X}_j$ as its column vectors. In order to reduce the disturbance between dictionary atoms associated to different classes, we proposed a latent-value weighted dictionary coherence term,

$$\sum\nolimits_{j=1}^{C}\sum\nolimits_{l\neq j}\sum\nolimits_{n=1}^{N}\sum\nolimits_{m\neq n}w_{j,m}\left(\boldsymbol{d}_m^T\boldsymbol{d}_n\right)^2 w_{l,n}$$

to promote the incoherence between dictionary atoms. If the dictionary atom $\boldsymbol{d}_m$ and $\boldsymbol{d}_n$ are very similar (i.e., the absolute value of $\boldsymbol{d}_m^T\boldsymbol{d}_n$ is big), the desired latent values, $w_{j,m}\times w_{l,n}$, will become smaller via minimizing the proposed weighted dictionary coherence term. Thus $\boldsymbol{d}_m$ and $\boldsymbol{d}_n$ would be more likely associated to the same class.

The latent matrix $\boldsymbol{W}$ and the learnt dictionary $\boldsymbol{D}$ do the latent representation together. When $w_{j,m}$ for $\boldsymbol{d}_m$ is large, under the sparse constraint of $\boldsymbol{X}_j$, $\boldsymbol{d}_m$ would more likely have a big contribution in the representation of $\boldsymbol{X}_j$, and then in the dictionary updating $\boldsymbol{X}_j$ would also have a bigger effect on the updating of $\boldsymbol{d}_m$.

3.2. Discussion of latent matrix

For a dictionary atom $\boldsymbol{d}_m$, let $\boldsymbol{v}_m$=[$w_{1,m}$,$w_{2,m}$,…,$w_{C,m}$]$\in\Re^{1\times C}$ be its latent values for all classes. This latent vector builds the relationship between $\boldsymbol{d}_m$ and all class labels. The constraints on latent value in Eq. (1) allow $\boldsymbol{d}_m$ represent different class data more flexibly than the class-specific dictionary atom and the shared dictionary atom. Both the class-specific dictionary learning and shared dictionary learning could be regarded as special cases of the proposed latent dictionary learning.

When all $\boldsymbol{v}_m$ have only one non-zero element (e.g., 1), each dictionary atom can only represent the data of a single class. In this case, the latent dictionary learning becomes discriminative class-specific dictionary learning

$$\begin{aligned}\min_{\boldsymbol{D},\boldsymbol{X}}&\sum\nolimits_{j=1}^{C}\left\|\boldsymbol{A}_j-\boldsymbol{D}_{(j)}\boldsymbol{X}_j\right\|_F^2+\lambda_1\left\|\boldsymbol{X}_j\right\|_1+\lambda_2\left\|\boldsymbol{X}_j-\boldsymbol{M}_j\right\|_F^2\\&+\lambda_3\sum\nolimits_{j=1}^{C}\sum\nolimits_{l\neq j}\left\|\boldsymbol{D}_{(j)}^T\boldsymbol{D}_{(l)}\right\|\end{aligned} \tag{2}$$

where $\boldsymbol{D}$=[$\boldsymbol{D}_{(1)}$, $\boldsymbol{D}_{(2)}$,…, $\boldsymbol{D}_{(C)}$], $\boldsymbol{D}_{(j)}$ is the sub-dictionary associated to the j-th class.

When all elements of $\boldsymbol{v}_m$ have the same value for every m, e.g., $w_{j,m}$=1 for each j and each m, each dictionary atom can represent the data of all classes. In this case, the latent dictionary learning changes to discriminative shared dictionary learning

$$\min_{\boldsymbol{D},\boldsymbol{X}} \sum_{j=1}^{C} \left\| \boldsymbol{A}_j - \boldsymbol{D}\boldsymbol{X}_j \right\|_F^2 + \lambda_1 \left\| \boldsymbol{X}_j \right\|_1 + \lambda_2 \left\| \boldsymbol{X}_j - \boldsymbol{M}_j \right\|_F^2 + \lambda_3 \sum_{n=1}^{N} \sum_{m \neq n} \left(\boldsymbol{d}_m^T \boldsymbol{d}_n \right)^2 \tag{3}$$

3.3. Latent classification model

After latent dictionary learning, the dictionary $\boldsymbol{D}$ and the latent matrix $\boldsymbol{W}$ are known. The latent vector $\boldsymbol{w}_j$ indicates the relationship between all atoms of $\boldsymbol{D}$ and j-th class. Since the latent value is non-negative, the latent vector, denoted by $\boldsymbol{\mu}=\sum_{j=1}^{C}\boldsymbol{w}_j$, reflects the total relationship between all atoms of $\boldsymbol{D}$ and all involved classes. A big value of μ_m shows dictionary atom $\boldsymbol{d}_m$ is important to the representation of all classes.

In the testing phase, for a testing sample $\boldsymbol{y}$ there are two coding strategies: globally coding $\boldsymbol{y}$ on the whole latent dictionary and locally coding $\boldsymbol{y}$ on the latent sub-dictionary associated to some class. Based on the learned latent dictionary $\boldsymbol{D}$ and $\boldsymbol{W}$, we proposed a latent-value weighted classification model.

When the training samples for each class are rather enough, the testing sample $\boldsymbol{y}$ is locally coded as (take j-th class as an example)

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\| \boldsymbol{y} - \boldsymbol{D}\text{diag}(\boldsymbol{w}_j)\boldsymbol{\alpha} \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \right\|_1 \tag{4}$$

Then the classification is conducted via

$$\text{identity}\left(\boldsymbol{y}\right) = \arg\min_j \left\{ \left\| \boldsymbol{y} - \boldsymbol{D}\text{diag}(\boldsymbol{w}_j)\hat{\boldsymbol{\alpha}} \right\|_2^2 + \lambda \left\| \hat{\boldsymbol{\alpha}} \right\|_1 \right\} \tag{5}$$

When the training samples for each class are not enough (e.g., in face recognition, action recognition), the testing sample $\boldsymbol{y}$ is globally coded as

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\| \boldsymbol{y} - \boldsymbol{D}\text{diag}(\boldsymbol{\mu})\boldsymbol{\alpha} \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \right\|_1 \tag{6}$$

Then the classification is conducted via

$$\text{identity}\left(\boldsymbol{y}\right) = \arg\min_j \left\{ \left\| \boldsymbol{y} - \boldsymbol{D}\text{diag}(\boldsymbol{w}_j)\hat{\boldsymbol{\alpha}} \right\|_2 \right\} \tag{7}$$

4. Optimization of LDL

The LDL objective function in Eq. (1) can be divided into two sub-problems by learning dictionary and latent matrix alternatively: updating $\boldsymbol{X}$ and $\boldsymbol{D}$ by fixing $\boldsymbol{W}$, and updating $\boldsymbol{W}$ by fixing $\boldsymbol{X}$ and $\boldsymbol{D}$.

4.1. Dictionary Learning

By fixing the latent matrix $\boldsymbol{W}$, the LDL model becomes

$$\min_{\boldsymbol{D},\boldsymbol{X}} \sum_{j=1}^{C} \left\| \boldsymbol{A}_j - \boldsymbol{D}\text{diag}\left(\boldsymbol{w}_j\right)\boldsymbol{X}_j \right\|_F^2 + \lambda_1 \left\| \boldsymbol{X}_j \right\|_1 + \lambda_2 \left\| \boldsymbol{X}_j - \boldsymbol{M}_j \right\|_F^2 + \lambda_3 \sum_{j=1}^{C} \sum_{l \neq j} \sum_{n=1}^{N} \sum_{m \neq n} w_{j,m} \left(\boldsymbol{d}_m^T \boldsymbol{d}_n \right)^2 w_{l,n} \tag{8}$$

The dictionary learning could also be optimized by alternatively solving $\boldsymbol{X}$ and $\boldsymbol{D}$. When $\boldsymbol{D}$ is fixed, the solving of $\boldsymbol{X}$ is a sparse coding problem with an additional within-class scatter term, which could be solved class by class:

$$\min_{\boldsymbol{X}_j} \left\| \boldsymbol{A}_j - \boldsymbol{D}\text{diag}\left(\boldsymbol{w}_j\right)\boldsymbol{X}_j \right\|_F^2 + \lambda_1 \left\| \boldsymbol{X}_j \right\|_1 + \lambda_2 \left\| \boldsymbol{X}_j - \boldsymbol{M}_j \right\|_F^2 \tag{9}$$

The problem could be solved efficiently by using the Iterative Projection Method [29] as [21].

When $\boldsymbol{X}$ is fixed, denote by $\boldsymbol{Y}=[\boldsymbol{Y}_1,\boldsymbol{Y}_2,\ldots,\boldsymbol{Y}_C]$ the latent coding coefficient, where $\boldsymbol{Y}_j=\text{diag}(\boldsymbol{w}_j)\boldsymbol{X}_j$, the dictionary learning problem of Eq. (8) changes to

$$\min_{\boldsymbol{D}} \left\| \boldsymbol{A} - \boldsymbol{D}\boldsymbol{Y} \right\|_F^2 + \lambda_3 \sum_{j=1}^{C} \sum_{l \neq j} \sum_{n=1}^{N} \sum_{m \neq n} w_{j,m} \left(\boldsymbol{d}_m^T \boldsymbol{d}_n \right)^2 w_{l,n} \tag{10}$$

Here we update the dictionary atom by atom. Let $\boldsymbol{\varGamma}=\boldsymbol{Y}\boldsymbol{Y}^T$, $\boldsymbol{\varLambda}=\boldsymbol{A}\boldsymbol{Y}^T$. For the updating of n-th dictionary atom, the object function could be rewritten as

$$\min_{\boldsymbol{d}_n} Tr\left(\boldsymbol{D}^T \boldsymbol{D} \boldsymbol{\varGamma} - 2\boldsymbol{D}^T \boldsymbol{\varLambda} \right) + 2\lambda_3 \sum_{m \neq n} \left(\boldsymbol{d}_m^T \boldsymbol{d}_n \right)^2 \sum_{j=1}^{C} \sum_{l \neq j} w_{j,m} w_{l,n} \tag{11}$$

Based on the dictionary updating algorithm of [30], $\boldsymbol{d}_n$ is updated by Eq. (12) and Eq. (13)

$$\boldsymbol{u} = \left(\varGamma_{n,n} \boldsymbol{I} + 2\lambda_3 \boldsymbol{I}_{\boldsymbol{Q}} \right)^{-1} \left(\boldsymbol{\varLambda}_n - \boldsymbol{D}\boldsymbol{\varGamma}_n - 2\lambda_3 \boldsymbol{Q}\boldsymbol{D}_n \right) \tag{12}$$

$$\boldsymbol{D}_n = \boldsymbol{d}_n = \boldsymbol{u} / \left\| \boldsymbol{u} \right\|_2 \tag{13}$$

where $\boldsymbol{Q}=\sum_{m\neq n}\boldsymbol{d}_m\boldsymbol{d}_m^T\sum_{j=1}^{C}\sum_{l\neq j}w_{j,m}w_{l,n}$, $\boldsymbol{I}$ is an identity matrix, $\boldsymbol{I}_{\boldsymbol{Q}}$ is a diagonal matrix with the same diagonal elements as $\boldsymbol{Q}$, $\varGamma_{n,n}$ is the element in n-th row and n-th column of $\boldsymbol{\varGamma}$, $\boldsymbol{\varGamma}_n$, $\boldsymbol{\varLambda}_n$, and $\boldsymbol{D}_n$ are the n-th column vectors of $\boldsymbol{\varGamma}$, $\boldsymbol{\varLambda}$, and $\boldsymbol{D}$, respectively. Eq. (13) normalizes each dictionary atom to have unit l_2-norm.

With Eq. (12) and Eq. (13), each atom of $\boldsymbol{D}$ could be updated. As described in [30], after several iterations the updating of $\boldsymbol{D}$ will converge.

4.2. Latent matrix learning

When $\boldsymbol{D}$ and $\boldsymbol{X}$ are learnt, we fix them and learn the latent matrix $\boldsymbol{W}$. Due to the constraint of $\sum_n w_{k,n}=\delta$ for k-th class, we update the latent matrix class by class. For the updating of $\boldsymbol{w}_k$ for k-th class, the LDL model of Eq. (1) changes to

$$\begin{aligned} &\min_{\boldsymbol{w}_k} \left\| \boldsymbol{A}_k - \boldsymbol{D}\text{diag}\left(\boldsymbol{w}_k\right)\boldsymbol{X}_k \right\|_F^2 \\ &\quad + 2\lambda_3 \sum_{n=1}^{N} w_{k,n} \sum_{m \neq n} \left(\boldsymbol{d}_m^T \boldsymbol{d}_n \right)^2 \sum_{j \neq k}^{C} w_{j,m} \\ &\quad \text{s.t. } w_{k,n} \geq 0 \ \forall\, n; \qquad \sum_n w_{k,n} = \delta \end{aligned} \tag{14}$$

which is a constrained quadratic programming problem. Here we proposed an efficient solver for Eq. (14).

Denote by $\boldsymbol{X}_{k,n}$ the n-th column vector of $\boldsymbol{X}_k$. Let $\boldsymbol{b}=[b_1,b_2,\ldots,b_N]^T$, $b_n=\sum_{m\neq n}(\boldsymbol{d}_m^T\boldsymbol{d}_n)^2\sum_{j\neq k}^{C}w_{j,m}$, $\boldsymbol{a}=\text{vec}(\boldsymbol{A}_k)$, $\boldsymbol{B}_n=\boldsymbol{d}_n\boldsymbol{X}_{k,n}$, and $\boldsymbol{R}=[\text{vec}(\boldsymbol{B}_1), \text{vec}(\boldsymbol{B}_2),\ldots,\text{vec}(\boldsymbol{B}_N)]$, where $\text{vec}(\boldsymbol{B})$ is a column vector generated from a matrix $\boldsymbol{B}$ by concatenating all column vectors of $\boldsymbol{B}$. Then the latent matrix learning problem could be rewritten as

$$\min \left\| \boldsymbol{a} - \boldsymbol{R}\boldsymbol{w}_k \right\|_F^2 + 2\lambda_3 \boldsymbol{b}^T \boldsymbol{w}_k \ \text{ s.t. } w_{k,n} \geq 0 \ \forall\, n; \sum_n w_{k,n} = \delta \tag{15}$$

Based on the framework of Iterative Projection Method (IPM) [29], Eq. (15) could be efficiently solved in an iterative procedure, where in t+1-th iteration we update the latent vector $\boldsymbol{w}_k^{(t+1)}$ as:

$$\boldsymbol{\tau}_0 = \boldsymbol{w}_k^{(t)} - \left(\boldsymbol{R}^T\left(\boldsymbol{R}\boldsymbol{w}_k^{(t)} - \boldsymbol{a}\right) + \lambda_3 \boldsymbol{b}\right) \Big/ \sigma \tag{16}$$

$$\boldsymbol{\tau}_1 = \boldsymbol{\tau}_0 - \left(\sum\nolimits_n \tau_{0,n} - \delta\right) \Big/ N \tag{17}$$

$$\boldsymbol{\tau}_2 = \max\left(\boldsymbol{\tau}_1, 0\right) \tag{18}$$

$$\boldsymbol{w}_k^{(t+1)} = \boldsymbol{\tau}_2 \Big/ \sum\nolimits_n \tau_{2,n} \cdot \delta \tag{19}$$

where σ is a scalar variable to control the step length of IPM, and max($\boldsymbol{\tau}$,0) is an operator to change the negative elements of $\boldsymbol{\tau}$ as 0. The detailed derivation of the solution of Eq. (15) is presented in the Appendix.

Table 1: Algorithm of Latent Dictionary Learning.

```
Latent Dictionary Learning (LDL)
1.  Initialization the latent matrix $W$
2.  Dictionary Learning
    While not converge do
        Update $X$ with fixed $D$.via solving Eq. (9).
        Update dictionary $D$ atom by atom by solving Eq. (10).
    End while
3.  Latent Matrix Learning
    Update the latent matrix $W$ via solving Eq. (14).
4.  Output
    Return to step 2 until the values of the objective function in Eq. (1) in adjacent iterations are close enough or the maximum number of iterations is reached.
    Output $D$ and $W$.
```

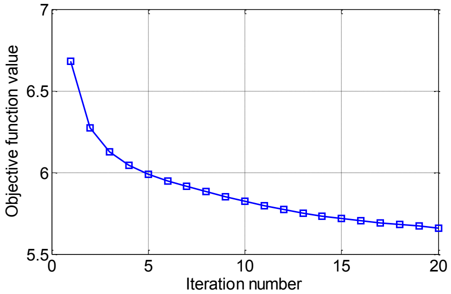

Figure 2: An example of LDL minimization process on UCF sport action dataset [31].

The whole algorithm of the proposed latent dictionary learning is summarized in Table 1. The algorithm will converge since the total cost function in Eq. (1) will decrease in two alternative optimizations. Fig.2 shows an example of LDL minimization on UCF sports action dataset [31].

5. Experiment results

In this section, we verify the performance of LDL on various classification tasks. Action classification, gender classification, and face recognition are performed by using LDL and the competing methods in Section 5.1, Section 5.2, and Section 5.3, respectively. To clearly illustrate the advantage of LDL, we compare LDL with several latest DL methods, such as Discriminative K-SVD (DKSVD) [10], Label Consistent K-SVD (LCKSVD) [11], dictionary learning with structure incoherence (DLSI) [20], dictionary learning with commonality and particularity (COPAR) [12], joint dictionary learning (JDL) [13] and Fisher discrimination dictionary learning (FDDL) [21]. Besides, we also report sparse representation based classifier (SRC) [5], linear support vector machine (SVM) and some methods for special tasks. In no specific instruction, the number of class-specific atoms in these DL methods is set as the number of training samples in the same class.

As described in Section 3.2, the latent vector $\boldsymbol{v}_m$ indicates the relationship between the m-th dictionary atom, $\boldsymbol{d}_m$, and class labels. In the initialization of $\boldsymbol{W}$=[$\boldsymbol{v}_1$;…;$\boldsymbol{v}_m$,…,$\boldsymbol{v}_N$], $\boldsymbol{v}_m$ is initialized to have only one non-zero element (e.g., 1). So latent dictionary is initialized by using class-specific atoms, of which the number is set as the number of training samples. In latent dictionary learning, the dictionary atom would be removed if its weight for all classes is less than a small positive scalar. So the number of latent dictionary atoms would be less than or equal to the initial number.

In all experiments, the parameters λ_1 and λ_2 in the dictionary learning phase and λ in the testing phase are determined via cross-validation. The parameter λ_3, which control the latent-value weighted dictionary coherence, is empirically set to 0.0001*P/(N*(N-1)), where P is the number of all training samples. For action classification and face recognition, global coding on the learned dictionary (i.e., Eq.(4)) is adopted (e.g., LDL-GC, FDDL-GC), while in gender classification, we report the classification results of Eqs. (4) and (6) (e.g., LDL-LC(GC), FDDL-LC(GC)).

5.1. Action Classification

The benchmark action dataset, UCF sports action dataset [31], is used to conduct the action classification experiment. The UCF sports action dataset [31] collected video clips from various broadcast sports channels (e.g., BBC and ESPN). The action bank features of 140 videos provided by [32] are adopted in the experiment. These videos cover 10 sport action classes: driving, golfing, kicking, lifting, horse riding, running, skateboarding, swinging-(prommel horse and floor), swinging-(high bar) and walking, some of which are shown in Fig. 3.

As the experiment setting in [14] and [11], we evaluated the LDL via five-fold cross validation. Here λ_3 is set to 0.1*P/(N*(N-1)), the dimension of the action bank feature is reduced to 100 via PCA, and the performance of some specific methods for action recognition, such as Qiu 2011 [14], action back feature with SVM classifier (Sadanand

2012) [32] are also reported. The recognition rates are listed in Table 2. We can observe that the proposed LDL achieves 95.0% accuracy, 1.4% improvement over the second best method, FDDL. With action bank features, all methods except DKSVD have the classification rates over 90%.

Following the leave-one-video-out experiment setting in [32], the proposed LDL could achieve 95.7% recognition accuracy, better than the state-of-the-art action recognition result (e.g., 95.0%) reported in [32].

Figure 3: Some video frames of the UCF sports action dataset.

Table 2: Classification rates (%) on the UCF sports action dataset.

Methods	Accuracy (%)	Methods	Accuracy (%)
Qiu 2011	83.6	LCKSVD	91.2
Sadanand 2012	90.7	COPAR	90.7
SRC	92.9	JDL	90.0
DLSI	92.1	FDDL	93.6
DKSVD	88.1	LDL	**95.0**

5.2. Gender Classification

a) AR database: As in [21], we chose a non-occluded subset (14 images per subject) from the AR face database [23], which consists of 50 males and 50 females, to conduct experiments of gender classification. Images of the first 25 males and 25 females were used for training, and the remaining 25 males and 25 females were used for testing. PCA was used to reduce the dimension of each image to 300.

The number training samples per class in gender classification is usually very large, so in gender classification we initially set the number of class-specific dictionary atoms in DL methods as 250. Table 3 lists the classification rates of all the competing methods. We can clearly see that LDL and other DL methods including class-specific dictionary atoms significantly outperform the shared dictionary learning methods, such as DKSVD and LCKSVD. The hybrid dictionaries (e.g., COPAR and JDL) are not better than class-specific dictionaries (e.g., FDDL and DLSI), which indicates that shared dictionary atoms are not powerful for classification. Our latent dictionary learning method with local-coding classifier (LDL-LC) has at least 1% improvement over all the other methods.

Class-specific dictionary usually has a big size. Here we reduce the number of class-specific dictionary atoms per class from 250 to 25, and then report the performance of LDL, JDL, COPAR, DLSI and FDDL in Table 4 (DKSVD and LCKSV are excluded since they don’t contain class-specific dictionary atoms). It can be seen that the accuracies of all methods drop a little. However, LDL-LC can still achieve 95.0% accuracy, much better than other methods. The latent matrix in LDL allows the learned dictionary atoms to more flexibly represent the data of different classes. Especially in the case that only a small number of dictionary atoms available, the latent matrix could involve more dictionary atoms to represent the data of a certain class.

Table 3: The gender classification rates (%) on the AR database.

Methods	Accuracy (%)	Methods	Accuracy (%)
SRC	93.0	COPAR	93.4
DLSI	94.0	JDL	92.6
DKSVD	86.1	FDDL-LC	94.3
LCKSVD	86.8	FDDL-GC	94.3
SVM	92.4	LDL-LC(GC)	**95.3** (**94.8**)

Table 4: The gender classification rates (%) on the AR database with 25 initialized class-specific dictionary atoms per class.

DLSI	COPAR	JDL	FDDL-LC(GC)	LDL-LC(GC)
93.7	93.0	91.0	93.7(92.1)	**95.0**(92.4)

Male samples

Female samples

Figure 4: Some samples of males and females from FRGC 2.0.

Table 5: The gender classification rates (%) on the FRGC 2.0 database.

Methods	Accuracy (%)	Methods	Accuracy (%)
SRC	93.0	COPAR	93.4
DLSI	94.5	JDL	90.8
DKSVD	85.6	S(U)-SC	94.7(93.2)
LCKSVD	89.5	FDDL-LC(GC)	**95.7**(94.1)
CNN	94.1	LDL-LC(GC)	**95.7**(94.6)

b) FRGC 2.0: We then conduct gender classification on the large-scale FRGC 2.0 database [33] with the same experiment setting as that in [34] and [35]. There are 568 individuals (243 females and 325 males) and 14,714 face images collected under various lighting conditions and backgrounds. Some samples of male and female are shown in Fig. 4. 3,014 images from randomly selected 114 subjects are used as the test set, with the rest 11,700 images as the training set. Here we use 300-dimensional PCA feature. The experimental results of DL methods are listed in Table 5, where the state-of-the-art S(U)-SC methods [35] and the CNN method [34] are also reported. One can see that LDL has similar performance with FDDL, while both LDL-LC and FDDL-LC are better than S(U)-SC, CNN, and other DL methods.

5.3. Face recognition

In this section, we evaluate LDL on the experiments of face recognition. We firstly conduct face recognition on a subset of FRGC 2.0 [33]. This subset collects the face images of 316 subjects from the query face dataset, where the selected subject should have no less than 10 samples. This subset contains 7318 faces images, which have large variation of lighting, accessory, expression, and image blur, etc. 5 samples per person are randomly chosen as training data, with the remaining images for testing. 300-d PCA features are used and all experiments were run 5 times to calculate the mean and standard deviation. The results of all competing methods are listed in Table 6. For fair comparison, we also report JDL and DLSI with global-coding classifier, denoted by JDL* and DLSI*. We can observe than LDL is slightly better than FDDL, and significantly outperform other methods. Both the original classifiers of DLSI and JDL locally encode the testing sample on the sub-dictionary of each class, which don't work well in face recognition. The hybrid DL models, COPAR and JDL, work worse than LDL. This may be because it is not easy to predefine some atoms as shared dictionary atoms with the remaining atoms as class-specific atoms. When the shared dictionary part is big, the discrimination of the hybrid dictionary would decrease.

Table 6: The face recognition rates (%) on the FRGC 2.0 database.

Methods	Accuracy (%)	Methods	Accuracy (%)
SRC	90.0±0.5	COPAR	89.6±0.5
DLSI	68.6±0.4	JDL	75.5±0.5
DLSI*	93.4±0.4	JDL*	91.2±0.5
DKSVD	79.6±0.5	FDDL	95.1±0.35
LCKSVD	80.2±0.8	LDL	**95.3±0.23**
SVM	72.9±0.7		

We also evaluate LDL on the application of face recognition in the wild. The aligned labeled face in the wild (LFWa) [25] is used here. LFW is a large-scale database, which contains variations of pose, illumination, expression, misalignment and occlusion, etc, as shown in Fig. 5. 143 subjects with no less than 11 samples per subject are chosen (4174 images in total). For each person the first 10 samples are used for training data with the remaining samples for testing. Histogram of Uniform-LBP is extracted via dividing a face image into 10×8 patches. Then we use PCA to reduce the histogram dimension to 1000. Table 7 illustrates the comparison of all methods. Similar to the results on FRGC, LDL achieves the best performance. Especially, the proposed LDL has over 4% improvement compared to the hybrid dictionary learning models.

Figure 5: Some face images of LFWa.

Table 7: The face recognition results of different methods on the LFW database.

Methods	Accuracy (%)	Methods	Accuracy (%)
SRC	72.7	COPAR	72.6
DLSI*	73.8	JDL*	72.8
DKSVD	65.9	FDDL	74.8
LCKSVD	66.0	LDL	**77.2**
SVM	63.0		

6. Conclusion

We proposed a latent dictionary learning (LDL) method, which learns a discriminative dictionary and a latent matrix jointly for sparse representation based classification. The latent matrix is learned to indicate the relationship between dictionary atoms and class labels. Instead of fixing the labels of dictionary atoms beforehand, the latent matrix is updated in the latent dictionary learning and the relationship between dictionary atoms and class labels is adaptively built. Meanwhile, the within-class term of coding coefficients and latent-value weighted dictionary coherence term ensure the latent dictionary to be discriminatively trained. The extensive experiments of classifying action, gender and face identity demonstrated the effectiveness of LDL to other latest dictionary learning based classification methods.

Acknowledgement

This work is partially supported by the National Natural Science Foundation of China under Grant no. 61272050

References

[1] M. Elad, and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE T. IP*, 15:(12):3736–3745, 2006.

[2] J.C. Yang, J. Wright, Y. Ma, and T. Huang. Image super-resolution as sparse representation of raw image patches. In *Proc. CVPR,* 2008.

[3] B.A. Olshausen, and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature,* 381: 607-609, 1996.

[4] B.A. Olshausen, and D.J. Field. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Research,* 37(23):3311-3325, 1997.

[5] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust Face Recognition via Sparse Representation. *IEEE T. PAMI*, 31(2):210–227, 2009.

[6] J.C. Yang, K. Yu, Y. Gong, and T. Huang, T. Linear spatial pyramid matching using sparse coding for image classification. In *Proc. CVPR*, 2009.

[7] K. Huang and S. Aviyente. Sparse representation for signal classification. In *Proc. NIPS*, 2006.

[8] J. Mairal, F. Bach, and J. Ponce. Task-Driven Dictionary Learning. *IEEE T. PAMI*, 34(4):791-804, 2012.

[9] M.Y. Zhou, H.J. Chen, J. Paisley, L. Ren, L.B. Li, Z.M. Xing, D. Dunson, G. Sapiro, and L. Carin. Nonparametric Bayesian Dictionary Learning for Analysis of Noisy and Incomplete Images. *IEEE T. IP*, 21(1):130-144, 2012.

[10] Q. Zhang and B.X. Li. Discriminative K-SVD for dictionary learning in face recognition. In *Proc. CVPR*, 2010.

[11] Z.L. Jiang, Z. Lin, and L.S. Davis. Label consistent K-SVD: Learning a discriminative dictionary for recognition. *IEEE Trans. PAMI*, 35(11): 2651-2664, 2013.

[12] S. Kong and D.H., Wang. A dictionary learning approach for classification: Separating the particularity and the commonality. In *Proc. ECCV*, 2012.

[13] N. Zhou and J.P. Fan. Learning inter-related visual dictionary for object recognition. In *Proc. CVPR*, 2012.

[14] Q. Qiu, Z.L. Jiang, and R. Chellappa. Sparse Dictionary-based Representation and Recognition of Action Attributes. In *Proc. ICCV*, 2011.

[15] T. Guha and R.K. Ward. Learning Sparse Representations for Human Action Recognition. *IEEE T. PAMI*, 34(8):1576-1888, 2012.

[16] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE T. SP*, 54(11):4311–4322, 2006.

[17] O. Bryt and M. Elad. Compression of facial images using the K-SVD algorithm. *Journal of Visual Communication and Image Representation*, 19(4):270–282, 2008.

[18] J.J. Wang, J.C. Yang, K. Yu, F.J. Lv, and T. Huang. Locality-constrained linear coding for image classification. In *Proc. CVPR*, 2010.

[19] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, A. Supervised dictionary learning. In *Proc. NIPS*, 2009.

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

[21] M. Yang, L. Zhang, X.C. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *Proc. ICCV*, 2011.

[22] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zissserman. Learning discriminative dictionaries for local image analysis. In *Proc. CVPR*, 2008.

[23] A. Martinez and R. Benavente. The AR face database. CVC Tech. Report No. 24, 1998.

[24] S. Bengio, F. Pereira, Y. Singer, and D. Strelow. Group sparse coding. In *Proc. NIPS*, 2009.

[25] L. Wolf, T. Hassner and Y. Taigman. Similarity Scores based on Background Samples. In *Proc. ACCV*, 2009.

[26] A. Castrodad and G. Sapiro. Sparse modeling of human actions from motion imagery. *Int'l Journal of Computer Vision*, 100:1-15, 2012.

[27] L. Shen, S.H. Wang, G. Sun, S.Q. Jiang, and Q.M. Huang. Multi-level discriminative dictionary learning towards hierarchical visual categorization. In *Proc. CVPR*, 2013.

[28] R. Rubinstein, A.M. Bruckstein, and M. Elad. Dictionaries for Sparse Representation Modeling, 98(6): 1045-1057, *Proceedings of the IEEE*, 2010.

[29] L. Rosasco, A. Verri, M. Santoro, S. Mosci, and S. Villa. Iterative Projection Methods for Structured Sparsity Regularization. MIT Technical Reports, MIT-CSAIL-TR-2009-050, CBCL-282.

[30] J. Marial, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11: 19-60, 2010.

[31] M. Rodriguez, J. Ahmed, and M. Shah. A spatio-temporal maximum average correlation height filter for action recognition. In *Proc. CVPR*, 2008.

[32] S. Sadanand and J.J. Corso. Action bank: A high-level representation of activeity in video. In *Proc. CVPR*, 2012.

[33] P.J. Phillips, P.J. Flynn, W.T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W.J. Worek. Overview of the face recognition grand challenge. In *Proc. CVPR*, 2005.

[34] K. Yu, W. Xu, and Y. Gong. Deep learning with kernel regularization for visual recognition. In *Proc. NIPS*, 2009.

[35] J.C. Yang, K. Yu, and T. Huang. Supervised Translation-Invariant Sparse coding. In *Proc. CVPR*, 2010.

[36] M. Yang, L. Zhang, X.C. Feng, and D. Zhang. Sparse representation based Fisher discrimination dictionary learning for image classification. To appear in *IJCV*.

Appendix: the solution of Eq. (15)

Based on the framework of IPM [29], Eq. (15) could be solved iteratively. In *t*+1-th iteration, the sub-problem of Eq. (15) is

$$\begin{aligned}&\min_{\boldsymbol{w}_k^{(t+1)}} \left\| \boldsymbol{w}_k^{(t+1)} - \boldsymbol{w}_k^{(t)} + \left(\boldsymbol{R}^T \left(\boldsymbol{R}\boldsymbol{w}_k^{(t)} - \boldsymbol{a} \right) + \lambda_3 \boldsymbol{b} \right) \big/ \sigma \right\|_2^2 \\ &\text{s.t. } w_{k,n}^{(t+1)} \geq 0 \ \ \forall n; \qquad \textstyle\sum_n w_{k,n}^{(t+1)} = \delta\end{aligned} \tag{20}$$

which is a special case of the following problem:

$$\min_{\boldsymbol{x}} \|\boldsymbol{\alpha} - \boldsymbol{x}\|_2^2 \text{ s.t. } x_n \geq 0 \ \ \forall n; \ \textstyle\sum_{n=1}^{N} x_n = \delta \tag{21}$$

Denote $\hat{\boldsymbol{\alpha}}$ the projection of $\boldsymbol{\alpha}$ onto the super-plane of $\sum_n x_n = \delta$. It could be derived that

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} - 1/N \cdot \left(\textstyle\sum_n \alpha_n - \delta \right) \tag{22}$$

Now the problem of Eq. (22) is equal to

$$\min_{\boldsymbol{x}} \|\hat{\boldsymbol{\alpha}} - \boldsymbol{x}\|_2^2 \text{ s.t. } x_n \geq 0 \ \ \forall n; \textstyle\sum_{n=1}^{N} x_n = \delta \tag{23}$$

This problem has an analytical solution, which is

$$\bar{\boldsymbol{\alpha}} = \max\left(\hat{\boldsymbol{\alpha}}, 0 \right) \tag{24}$$

$$\hat{\boldsymbol{x}} = \pi \bar{\boldsymbol{\alpha}} \big/ \textstyle\sum_n \bar{\alpha}_n \tag{25}$$