Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Learning Sparse Representations from Datasets with Uncertain Group Structures: Model, Algorithm and Applications

Longwen Gao and **Shuigeng Zhou**
Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science
Fudan University, Shanghai 200433, China
{lwgao, sgzhou}@fudan.edu.cn

Abstract

Group sparsity has drawn much attention in machine learning. However, existing work can handle only datasets with *certain* group structures, where each sample has a *certain* membership with one or more groups. This paper investigates the learning of sparse representations from datasets with *uncertain* group structures, where each sample has an *uncertain* membership with all groups in terms of a *probability distribution*. We call this problem *uncertain group sparse representation* (UGSR in short), which is a generalization of the standard *group sparse representation* (GSR). We formulate the UGSR model and propose an efficient algorithm to solve this problem. We apply UGSR to text emotion classification and aging face recognition. Experiments show that UGSR outperforms standard *sparse representation* (SR) and standard GSR as well as fuzzy kNN classification.

Introduction

In many regression problems and their applications to machine leaning and signal processing, regularization by sparsity-inducing norms has drawn a lot of research interest. For example, in the ANOVA problem, important main effects and interactions are often selected for accurate prediction (Yuan and Lin 2006). *Sparse representation* (SR) (Olshausen and Field 1997) using ℓ_1-norm selects a few relevant support signals and has some theoretical advantages (Hoyer 2003) in signal processing. For applications that require some definite sparsity patterns (Jenatton, Audibert, and Bach 2011), regularizers of structured sparsity were introduced for a better adaption to various tasks. For example, *group sparse representation* (GSR) (Yuan and Lin 2006) imposes the sparsity among groups of signals, composite absolute penalties (CAPs) (Zhao, Rocha, and Yu 2006) put a hierarchical group structure among signals. More complicate structures such as overlapping groups and graph structures were also proposed (Jacob, Obozinski, and Vert 2009).

Essentially, the above work as in (Zhao, Rocha, and Yu 2006) addresses two major concerns: how do the groups relate to each other? and how do the *samples* (or *signals* in signal processing context) within each group relate to each other? The former characterizes the sparsity among groups, which is measured by a sparse-inducing norm such as ℓ_1-norm, while the latter characterizes the sample "concentration" of groups, which is measured by an ℓ_γ-norm ($\gamma > 1$). However, in our point of view, there is a third concern that does not receive enough attention and thus needs further investigation: how do the samples relate to the groups? Existing work assumes that samples have *certain* membership with one or more groups. Concretely, early group sparsity models specify that each sample belongs to only one *certain* group. (Jacob, Obozinski, and Vert 2009) handles overlapping group structures that allow one sample to belong to several groups, by dividing the corresponding coefficient of a sample into several parts so that each of the overlapping groups has one part of the coefficient. Though overlapping groups extend the sample-group relationship from *many-to-one* to *many-to-many*, the sample-group relationship is still assumed to be *certain*. Recently, (Chen et al. 2013) deals with the problem of learning dictionaries from ambiguously labeled data, where training samples have multiple given labels, but only one is correct. Based on the dataset with noisy yet certain sample-group relationships, the authors tried to learn a dictionary with correct labeling by iterating between updating a confidence matrix and learning a dictionary from the clusters inferred by the confidence matrix. When this process converges, sparse representation with certain group structure can be performed on the learned dictionary. However, in reality, sample-group relationship can be *uncertain*. For example, in text classification, each training text may be probabilistically related to multiple classes; and in image understanding, a picture may be annotated to several categories based on a probability scheme.

This paper addresses a more general problem that learns group sparsity from datasets with *uncertain group structures*, where each sample is related to all groups in terms of a *probability distribution*. We propose a novel structured sparse representation called *uncertain group sparse representation* (UGSR) to deal with uncertain group structures. UGSR is a generalization of standard GSR.

SR and GSR have been successfully applied to many classification tasks such as images (Majumdar and Ward 2009; Wright et al. 2009), texts (Sainath et al. 2010) and biological data (Li and Ngom 2012; Yuan et al. 2012). While dealing with classification tasks, GSR usually outperforms

SR (Majumdar and Ward 2009) because group sparsity works better when the underlying samples are strongly group-sparse (Huang and Zhang 2009). However, as we have mentioned above, standard SR and GSR cannot handle probabilistic labels in classification tasks. On the other hand, in the literature of classification, fuzzy classification deals with how to assign objects to different classes based on fuzzy set theory. For example, the fuzzy k-NN (FkNN) method employs fuzzy set theory to predict fuzzy class membership (Keller, Gray, and Givens 1985). However, such methods are used mainly for classification or clustering, and can handle only fuzzy class membership. On the contrary, UGSR is a generalization of the standard GSR and is applicable to all situations of sparse selection. Furthermore, our experiments show that UGSR outperforms standard SR and GSR as well as the fuzzy kNN method in text emotion classification and aging face recognition.

Contributions of this paper are as follows: 1) We propose a novel group sparse representation model UGSR that can handle uncertain group structures. 2) We show that UGSR is a generalization of the standard GSR in a higher dimensional affine space. 3) We define the classification rule for UGSR and apply UGSR to classification tasks with probabilistic or weighted labels. 4) We conduct experiments to validate the UGSR model and algorithm. Experimental results show that UGSR outperforms standard GSR and SR as well as the fuzzy kNN method.

Preliminaries

Here we briefly introduce standard SR and GSR in the context of classification. Assume that we have M training samples in $\mathcal{R}^d$ that fall into G different classes, each training sample i has a label in $\{1..G\}$. Given a test sample $\boldsymbol{y}$, we are to label it according to the labels of the training samples.

Sparse representation (SR) represents the test sample as a linear combination of the training samples while requiring the coefficients to be sparse. These training samples together constitute a dictionary $\boldsymbol{D} \in \mathcal{R}^{d\times M}$. SR is to solve the following optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathcal{R}^M} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2 + \lambda R_1(\boldsymbol{x}) \right\} \\ &with\ R_1(\boldsymbol{x}) = \sum_{i=1}^{M} |\boldsymbol{x}_i|, \end{aligned} \tag{1}$$

where $\lambda > 0$ is a tradeoff parameter. The first term is the regression error, and the second term R_1 is an ℓ_1-norm that imposes sparsity to the coefficient vector $\boldsymbol{x}$.

Group sparse representation (GSR) uses label information during representation by requiring the coefficients corresponding to different class labels to be sparse. Let $\mathcal{G}_g$ be the group of indices of training samples with label $g \in \{1..G\}$, GSR can be formulated as:

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathcal{R}^M} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2 + \lambda R_2(\boldsymbol{x}) \right\} \\ &with\ R_2(\boldsymbol{x}) = \sum_{g=1}^{G} \sqrt{\sum_{i\in\mathcal{G}_g} \boldsymbol{x}_i^2} = \sum_{g=1}^{G} \left\|\boldsymbol{x}_{\mathcal{G}_g}\right\|_2, \end{aligned} \tag{2}$$

where $\lambda > 0$ is a tradeoff parameter. The first term is the regression error as in SR, and the second term R_2 can be seen as an $\ell_{1/2}$-norm: the ℓ_2-norm is for the elements of the coefficient vector $\boldsymbol{x}$ inside each group and is used as an indicator of the "concentration" of samples, different from the group "sparsity" measured by the ℓ_1-norm.

After the coefficient vector $\boldsymbol{x}$ in Eq. (1) or (2) is computed, we can decide which is the most suitable label for a test sample $\boldsymbol{y}$. The maximum ℓ_2 support rule (Sainath et al. 2010) works well with both SR and GSR. It classifies a test sample $\boldsymbol{y}$ as follows:

$$label^* = \arg\max_{g\in\{1..G\}} \left\|\boldsymbol{x}_{\mathcal{G}_g}\right\|_2. \tag{3}$$

The UGSR Model

Uncertain group sparse representation (UGSR)

We define the *uncertain group structure* underlying a dataset as follows: given a dictionary $\boldsymbol{D}$=$[\boldsymbol{D}^1 \dots \boldsymbol{D}^M] \in \mathcal{R}^{d\times M}$, which is a collection of M sample vectors in $\mathcal{R}^d$, and $\boldsymbol{D}^i$ is the i-th sample. Those samples belong to G groups labeled by $1..G$. The uncertain group structure implies the probabilistic relationship between each sample and each group. Assume that the i-th sample associates with the g-th group by a given probability $\boldsymbol{P}_g^i$, where $g \in \{1..G\}$ and $i \in \{1..M\}$. Then, we denote $\boldsymbol{P}^i$ as the *probability distribution vector* of sample i with regard to all groups, and the g-th element of the distribution vector is probability $\boldsymbol{P}_g^i$. Since $\boldsymbol{P}^i$ is a distribution, we have $\sum_{g\in\{1..G\}} \boldsymbol{P}_g^i = 1$ for all i, and $\boldsymbol{P}_g^i \geq 0$ for all i and g. The distribution vectors of all samples in dictionary $\boldsymbol{D}$ form a *distribution matrix* $\boldsymbol{P}$=$[\boldsymbol{P}^1 \dots \boldsymbol{P}^M]$. Given dictionary $\boldsymbol{D}$ and the corresponding distribution matrix $\boldsymbol{P}$, for a new sample $\boldsymbol{y} \in \mathcal{R}^d$, *uncertain group sparse representation* (UGSR) is to represent $\boldsymbol{y}$ as a sparse linear combination of all vectors in $\boldsymbol{D}$ by using $\boldsymbol{P}$.

To handle uncertain group structures, we have to take into consideration the group distributions of samples. We define the concept of *group distribution sparsity* as follows:

Definition 1. *Given a group distribution $\boldsymbol{P}^i$ of sample i, its sparsity $Sp(\boldsymbol{P}^i)$ is defined as:*

$$Sp(\boldsymbol{P}^i) = \sqrt{\|\boldsymbol{P}^i\|_{\frac{1}{2}}} = \sum_{g=1}^{G} \sqrt{\boldsymbol{P}_g^i}. \tag{4}$$

We use the square root of $\ell_{\frac{1}{2}}$-norm to indicate the sparsity of $\boldsymbol{P}^i$ because ℓ_0-norm is too strict and ℓ_1-norm is a constant by the equality $\sum_{g\in\{1..G\}} \boldsymbol{P}_g^i$=1. So a sparser $\boldsymbol{P}^i$ means its component values concentrate on fewer groups. Such a distribution is more informative because its entropy is smaller according to information theory. In contrast, a non-sparse distribution is more uniform and thus may contribute less in reducing the number of representing vectors. We formulate the UGSR model as below:

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathcal{R}^M} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2 + \lambda R_p(\boldsymbol{x}) \right\} \\ &with\ R_p(\boldsymbol{x}) = \sum_{g=1}^{G} \sqrt{\sum_{i=1}^{M} \boldsymbol{P}_g^i \boldsymbol{x}_i^2}, \end{aligned} \tag{5}$$

where $\lambda > 0$ is a tradeoff parameter. The first term is the regression error that is the same as in SR and GSR. The second term $R_p(\boldsymbol{x})$ is the uncertain group sparse regularizer that ensures dual sparsity: the selected samples should be sparse and their group distributions should also be sparse. Note that R_p concerns about the *relationship between samples and groups*, which is different from the regularizer in (Jenatton, Audibert, and Bach 2011) that adds weights to balance the overlapping groups, *i.e.*, the latter concerns about the *relationship among groups*. We have the following theorems to elaborate the properties of the UGSR model.

Theorem 1. *$\forall \boldsymbol{x} \in \mathcal{R}^M$, $\exists C_1, C_p \in \mathcal{R}$ such that $R_1(\boldsymbol{x}) \leq C_1 R_p(\boldsymbol{x})$ and $R_p(\boldsymbol{x}) \leq C_p R_1(\boldsymbol{x})$ hold.*

Proof. This result is straightforward since it can be shown that R_p is also a norm defined in $\mathcal{R}^M$, and $\mathcal{R}^M$ is a finite dimensional Hilbert space. Thus, they are equivalent norms, and so the theorem is proved. □

Theorem 2. *$\forall C \in \mathcal{R}$, $\forall i, j \in \{1..M\}$, if $\boldsymbol{x}$ minimizes $R_p(\boldsymbol{x})$ with respect to $R_1(\boldsymbol{x}) = C$, and if $Sp(\boldsymbol{P}^i) < Sp(\boldsymbol{P}^j)$, then $|\boldsymbol{x}_i| > |\boldsymbol{x}_j|$.*

Proof. As we are to prove the relation between two absolute component values of $\boldsymbol{x}$, let $\boldsymbol{u}_i$=$|\boldsymbol{x}_i| \geq 0$ for i=1..M, we have

$$R_1(\boldsymbol{x}) = \sum_{i=1}^{M} \boldsymbol{u}_i, R_p(\boldsymbol{x}) = \sum_{g=1}^{G} \sqrt{\sum_{i=1}^{M} \boldsymbol{P}_g^i \boldsymbol{u}_i^2}. \quad (6)$$

Considering the fact that $\boldsymbol{u}$ minimizes $f(\boldsymbol{u}) = \sum_{g=1}^{G} \sqrt{\sum_{i=1}^{M} \boldsymbol{P}_g^i \boldsymbol{u}_i^2} + \alpha \left(\sum_{i=1}^{M} \boldsymbol{u}_i - C \right)$, we have

$$\frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{u}_i} = \sum_{g=1}^{G} \frac{\boldsymbol{P}_g^i \boldsymbol{u}_i}{\sqrt{\sum_{l=1}^{M} \boldsymbol{P}_g^l \boldsymbol{u}_l^2}} + \alpha = 0, \forall i \in \{1..M\}. \quad (7)$$

Thus,

$$-\alpha = \sum_{g=1}^{G} \frac{\boldsymbol{P}_g^i \boldsymbol{u}_i}{\sqrt{\sum_{l=1}^{M} \boldsymbol{P}_g^l \boldsymbol{u}_l^2}}, \forall i \in \{1..M\}. \quad (8)$$

Considering the indexes of two elements in $\boldsymbol{u}$: i and j, the following equation holds:

$$\sum_{g=1}^{G} \frac{\boldsymbol{P}_g^i \boldsymbol{u}_i}{\sqrt{\sum_{l=1}^{M} \boldsymbol{P}_g^l \boldsymbol{u}_l^2}} = \sum_{g=1}^{G} \frac{\boldsymbol{P}_g^j \boldsymbol{u}_j}{\sqrt{\sum_{l=1}^{M} \boldsymbol{P}_g^l \boldsymbol{u}_l^2}}. \quad (9)$$

Let us focus on the variables associated with i and j, and consider the rest variables fixed, this is the same as the case when M=2. So without loss of generalization, let G=2, M=2, i=1 and j=2, Eq. (9) becomes

$$\begin{aligned} &\frac{\boldsymbol{P}_1^1 \boldsymbol{u}_1}{\sqrt{\boldsymbol{P}_1^1 \boldsymbol{u}_1^2 + \boldsymbol{P}_1^2 \boldsymbol{u}_2^2}} + \frac{\boldsymbol{P}_2^1 \boldsymbol{u}_1}{\sqrt{\boldsymbol{P}_2^1 \boldsymbol{u}_1^2 + \boldsymbol{P}_2^2 \boldsymbol{u}_2^2}} \\ &= \frac{\boldsymbol{P}_1^2 \boldsymbol{u}_2}{\sqrt{\boldsymbol{P}_1^1 \boldsymbol{u}_1^2 + \boldsymbol{P}_1^2 \boldsymbol{u}_2^2}} + \frac{\boldsymbol{P}_2^2 \boldsymbol{u}_2}{\sqrt{\boldsymbol{P}_2^1 \boldsymbol{u}_1^2 + \boldsymbol{P}_2^2 \boldsymbol{u}_2^2}}. \end{aligned} \quad (10)$$

The left part is monotonic with respect to $\boldsymbol{u}_1$, which increases as $\boldsymbol{u}_1$ becomes larger, and so is the right part with respect to $\boldsymbol{u}_2$. By the distribution property, we have

$$\boldsymbol{P}_1^1 + \boldsymbol{P}_2^1 = \boldsymbol{P}_1^2 + \boldsymbol{P}_2^2 = 1. \quad (11)$$

By $Sp(\boldsymbol{P}^1) < Sp(\boldsymbol{P}^2)$ we have $\sqrt{\boldsymbol{P}_1^1} + \sqrt{\boldsymbol{P}_2^1} < \sqrt{\boldsymbol{P}_1^2} + \sqrt{\boldsymbol{P}_2^2}$. Square on both sides, we have $\boldsymbol{P}_1^1 \boldsymbol{P}_2^1 < \boldsymbol{P}_1^2 \boldsymbol{P}_2^2$. Combine this with Eq. (11), we get $\left| \boldsymbol{P}_1^1 - \boldsymbol{P}_2^1 \right| > \left| \boldsymbol{P}_1^2 - \boldsymbol{P}_2^2 \right|$. By a simple discussion on the signs inside the absolute formulas, we can get the following inequality:

$$\frac{\boldsymbol{P}_1^1 - \boldsymbol{P}_1^2}{\sqrt{\boldsymbol{P}_1^1 + \boldsymbol{P}_1^2}} < \frac{\boldsymbol{P}_2^2 - \boldsymbol{P}_2^1}{\sqrt{\boldsymbol{P}_2^1 + \boldsymbol{P}_2^2}}. \quad (12)$$

Let $\boldsymbol{u}_1$=$\boldsymbol{u}_2$ in Eq. (10), we have

$$\frac{\boldsymbol{P}_1^1 - \boldsymbol{P}_1^2}{\sqrt{\boldsymbol{P}_1^1 + \boldsymbol{P}_1^2}} = \frac{\boldsymbol{P}_2^2 - \boldsymbol{P}_2^1}{\sqrt{\boldsymbol{P}_2^1 + \boldsymbol{P}_2^2}}. \quad (13)$$

Considering Ineq. (12) and Eq. (13), to satisfy Eq. (10), we have to increase $\boldsymbol{u}_1$ and decrease $\boldsymbol{u}_2$. So we have $\boldsymbol{u}_1$>$\boldsymbol{u}_2$. □

Theorem 1 indicates that minimizing R_1 or R_p to a certain degree will cause the other one to decrease. Theorem 2 means that R_p prefers to assign larger values to samples with sparser group distributions. Thus for coefficient vectors achieve the same R_1, the R_p regularizer prefers the one whose elements correspond to samples with sparser group distributions, and consequently the coefficient vectors generated by UGSR are more informative and useful. The following theorem describes the relationship between the UGSR regularizer and the standard GSR regularizer.

Theorem 3. *Let R_p be an uncertain group sparse regularizer on $\mathcal{R}^M$ with G uncertain groups, there exists a group sparse regularizer R_2 on $\mathcal{R}^N$ ($N = GM$) such that $R_p = R_2 \circ \boldsymbol{B}$, where $\boldsymbol{B} = [\sqrt{\boldsymbol{B}_1}; \ldots; \sqrt{\boldsymbol{B}_G}] \in \mathcal{R}^{N \times M}$ and $\forall g \in \{1..G\}$, $\boldsymbol{B}_g \in \mathcal{R}^{M \times M}$ is a diagonal matrix with diagonal elements $\boldsymbol{B}_g(i,i) = \boldsymbol{P}_g^i$, $i \in \{1..M\}$.*

Proof. By the definitions of uncertain regularizer R_p and diagonal matrices $\boldsymbol{B}_g$, $g \in \{1..G\}$, R_p can be rewritten as

$$R_p(\boldsymbol{x}) = \sum_{g=1}^{G} \sqrt{\boldsymbol{x}^\top \boldsymbol{B}_g \boldsymbol{x}} = \sum_{g=1}^{G} \|\boldsymbol{x}\|_{\boldsymbol{B}_g}. \quad (14)$$

If we write those $\|\boldsymbol{x}\|_{\boldsymbol{B}_g}$ in the form of column vectors, the sum above can be seen as an ℓ_1-norm as below:

$$\sum_{g=1}^{G} \|\boldsymbol{x}\|_{\boldsymbol{B}_g} = \left\| \begin{pmatrix} \sqrt{\boldsymbol{x}^\top \boldsymbol{B}_1 \boldsymbol{x}} \\ \vdots \\ \sqrt{\boldsymbol{x}^\top \boldsymbol{B}_G \boldsymbol{x}} \end{pmatrix} \right\|_1. \quad (15)$$

Since each $\boldsymbol{B}_g$ is a diagonal matrix and its elements are all non-negative, we rewrite it as the product of its square root $\boldsymbol{B}_g = \sqrt{\boldsymbol{B}_g}^\top \sqrt{\boldsymbol{B}_g}$. Thus,

$$\sqrt{\boldsymbol{x}^\top \boldsymbol{B}_g \boldsymbol{x}} = \sqrt{\boldsymbol{x}^\top \sqrt{\boldsymbol{B}_g}^\top \sqrt{\boldsymbol{B}_g} \boldsymbol{x}} = \left\| \sqrt{\boldsymbol{B}_g} \boldsymbol{x} \right\|_2. \quad (16)$$

So the R_p regularizer in Eq. (15) can be seen as a group sparse regularizer formed by an ℓ_1/ℓ_2-norm:

$$\left\| \begin{pmatrix} \left\|\sqrt{\boldsymbol{B}_1}\boldsymbol{x}\right\|_2 \\ \vdots \\ \left\|\sqrt{\boldsymbol{B}_G}\boldsymbol{x}\right\|_2 \end{pmatrix} \right\|_1 = \left\| \begin{pmatrix} \sqrt{\boldsymbol{B}_1}\boldsymbol{x} \\ \vdots \\ \sqrt{\boldsymbol{B}_G}\boldsymbol{x} \end{pmatrix} \right\|_{1,2}. \quad (17)$$

If we define the group sparse regularizer R_2 as the ℓ_1/ℓ_2-norm on $\mathcal{R}^N$ (N=GM) and let $\boldsymbol{B}$=$[\sqrt{\boldsymbol{B}_1};\ldots;\sqrt{\boldsymbol{B}_G}] \in \mathcal{R}^{N\times M}$, R_p is actually a composition of the higher dimensional group sparse regularizer R_2 and the affine transformation B, *i.e.*, $R_p(\boldsymbol{x}) = R_2(\boldsymbol{B}\boldsymbol{x})$. For R_2, there are G groups, each of which contains M samples. □

It can be seen that the uncertain group sparse regularizer R_p and the group sparse regularizer R_2 are the same when $\boldsymbol{P}$ is assigned with binary values (each $\boldsymbol{P}^i$ has exactly one "1"). That is, for *certain* group structures, UGSR degenerates into GSR. Therefore, R_p is a generalized group sparse regularizer in a higher dimensional affine space.

Classification based on UGSR

We consider both *hard classification* and *soft classification*. Here probabilistic group distribution is used to describe group uncertainty of training samples. Given a test sample $\boldsymbol{y} \in \mathcal{R}^d$, *hard classification* is to predict $\boldsymbol{y}$'s most-likely class label, while *soft classification* is to compute $\boldsymbol{y}$'s probability distribution over all classes.

In classification using GSR, the classes of training samples are used as groups. In the uncertain group structure setting, each training sample relates to all groups, which makes the standard group sparse regularizer impose less sparsity. When applying the UGSR model, we actually provide dual sparsity to test samples: the sparsity of training samples as in SR and the sparsity of their group distributions. The second sparsity makes UGSR favorably select training samples that are more informative and useful for classification.

Hard classification. In classification using SR and GSR, the maximum ℓ_2 support rule is used to determine the class label of $\boldsymbol{y}$. However, in our uncertain setting, the ℓ_2 supports of different classes may be equal to each other as each sample is related to all classes, this may degrade the performance of the maximum ℓ_2 support classification rule. So we propose a generalized maximum ℓ_2 support rule for hard classification based on the UGSR model:

$$label^* = \arg\max_{g\in\{1..G\}} \left\{ \left\| \sqrt{\boldsymbol{B}_g}\boldsymbol{x} \right\|_2 \right\}, \quad (18)$$

where $\left\|\sqrt{\boldsymbol{B}_g}\boldsymbol{x}\right\|_2$ indicates sample "concentration" of group g, which is consistent with the same concept in standard group sparsity and the maximum ℓ_2 support rule in a higher dimensional affine space.

Soft classification. The goal is to compute the probability distribution of test sample $\boldsymbol{y}$ associating with all classes (or groups). For UGSR, we use the normalized value of $\left\|\sqrt{\boldsymbol{B}_g}\boldsymbol{x}\right\|_2$ to measure the probability $\boldsymbol{P}_g^*$ of $\boldsymbol{y}$ belonging to class g. So for each class $g \in \{1..G\}$, we have

$$\boldsymbol{P}_g^* = \frac{\left\|\sqrt{\boldsymbol{B}_g}\boldsymbol{x}\right\|_2}{\sum_{h=1}^{G}\left\{\left\|\sqrt{\boldsymbol{B}_h}\boldsymbol{x}\right\|_2\right\}}. \quad (19)$$

Algorithm 1 Proximity Algorithm for SR and GSR

Input: dictionary $\boldsymbol{D}$, test vector $\boldsymbol{y}$
Initialize $\boldsymbol{x} = \boldsymbol{0}$
repeat
  $\boldsymbol{u} = \boldsymbol{x} - \frac{1}{L}\boldsymbol{D}^\top(\boldsymbol{D}\boldsymbol{x} - \boldsymbol{y})$
  $\boldsymbol{x} = Prox_{\frac{\lambda}{L}R}(\boldsymbol{u})$
until $\boldsymbol{x}$ converges

For comparison, we also let SR and GSR output a probability distribution of test sample $\boldsymbol{y}$ associating with all classes. The probability $\boldsymbol{P}_g^*$ of $\boldsymbol{y}$ belonging to class g is evaluated by

$$\boldsymbol{P}_g^* = \frac{\left\|\boldsymbol{x}_{\mathcal{G}_g}\right\|_2}{\sum_{h=1}^{G}\left\|\boldsymbol{x}_{\mathcal{G}_h}\right\|_2}. \quad (20)$$

The UGSR Algorithm

Here we present an efficient algorithm to solve the UGSR model via proximity operator.

Basic proximal method

Let $f(\boldsymbol{x}) = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\right\|_2^2$ and R be any sparse-inducing regularizer we mentioned above. As f is a smooth differentiable function, we can linearize it around the current point $\boldsymbol{x}^t$ at each iteration and reformulate it as in (Bach et al. 2011), the optimization within the iteration becomes

$$\min_{\boldsymbol{x}\in\mathcal{R}^M}\left\{\frac{1}{2}\left\|\boldsymbol{x} - (\boldsymbol{x}^t - \frac{1}{L}\nabla f(\boldsymbol{x}^t))\right\|_2^2 + \frac{\lambda}{L}R(\boldsymbol{x})\right\}, \quad (21)$$

where L is the upper bound of $\left\|\boldsymbol{D}^\top\boldsymbol{D}\right\|_2$ (Argyriou et al. 2011). Here, we recall the definition of proximity operator introduced by Moreau (Moreau 1962) as follows:

Definition 2. *φ is a real-valued convex function on $\mathcal{R}^M$, its proximity operator $Prox_\varphi\boldsymbol{x}(\boldsymbol{u})$ is defined as:*

$$\arg\min_{\boldsymbol{x}}\left\{\frac{1}{2}\left\|\boldsymbol{u} - \boldsymbol{x}\right\|_2^2 + \varphi(\boldsymbol{x}) : \boldsymbol{u} \in \mathcal{R}^M\right\}. \quad (22)$$

Let $\boldsymbol{u}$ be the standard gradient descent of $\boldsymbol{x}^t$ and φ be the sparse-inducing regularizer R, so $\boldsymbol{u} = \boldsymbol{x}^t - \frac{1}{L}\nabla f(\boldsymbol{x}^t)$, the update rule for Eq. (21) is

$$\boldsymbol{x} = Prox_{\frac{\lambda}{L}R}(\boldsymbol{x}^t - \frac{1}{L}\nabla f(\boldsymbol{x}^t)). \quad (23)$$

This optimization process is shown in Algorithm 1, which is applied to solving the optimization problems of SR and GSR since $Prox_{\frac{\lambda}{L}R_1}(\boldsymbol{u})$ and $Prox_{\frac{\lambda}{L}R_2}(\boldsymbol{u})$ can be computed directly (Bach et al. 2011).

Proximity operator computation of the uncertain group sparse regularizer R_p

As we have shown, R_p can be reformulated as a composition of a group regularizer R_2 on $\mathcal{R}^N$ and an affine transformation $\boldsymbol{B}$ in Theorem 3. Combining Eq. (21), Eq. (22) and Theorem 3, we get the following optimization problem:

$$\min_{\boldsymbol{x}}\left\{\frac{1}{2}\left\|\boldsymbol{u} - \boldsymbol{x}\right\|_2^2 + \frac{\lambda}{L}R_2(\boldsymbol{B}\boldsymbol{x}) : \boldsymbol{u} \in \mathcal{R}^M\right\}. \quad (24)$$

Algorithm 2 Proximity Algorithm for UGSR

Input: dictionary $\boldsymbol{D}$, distribution matrix $\boldsymbol{P}$, test vector $\boldsymbol{y}$
Initialize $\boldsymbol{x} = \boldsymbol{0}$
$\boldsymbol{B} = [\sqrt{diag(\boldsymbol{P}_1)}; \ldots; \sqrt{diag(\boldsymbol{P}_G)}]$
repeat
  $\boldsymbol{u} = \boldsymbol{x} - \frac{1}{L}\boldsymbol{D}^\top(\boldsymbol{D}\boldsymbol{x} - \boldsymbol{y})$
  Initialize $\boldsymbol{v} = \boldsymbol{0}$
  repeat
    $\boldsymbol{w} = \boldsymbol{v} - c\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{v} + \boldsymbol{B}\boldsymbol{u}$
    $\boldsymbol{v} = \kappa\boldsymbol{v} + (1-\kappa)\left(\boldsymbol{w} - Prox_{\frac{\lambda}{cL}R_2}(\boldsymbol{w})\right)$
  until $\boldsymbol{v}$ converges
  $\boldsymbol{x} = \boldsymbol{u} - c\boldsymbol{B}^\top\boldsymbol{v}$
until $\boldsymbol{x}$ converges

R_2 can be computed directly as in (Bach et al. 2011). Let $\mathcal{G}_g$=$\{(g-1)M$+1,$\cdots$,$gM\}$ be a group of indices and $\boldsymbol{w} \in \mathcal{R}^N$. Then,

$$[Prox_{\mu R_2}(\boldsymbol{w})]_{\mathcal{G}_g} = \left(1 - \frac{\mu}{\left\|\boldsymbol{w}_{\mathcal{G}_g}\right\|_2}\right)_+ \boldsymbol{w}_{\mathcal{G}_g} \tag{25}$$

for $g \in \{1..G\}$, where $\mu > 0$ and $(\cdot)_+ = \max(\cdot, 0)$. Now we adapt the theorem from (Micchelli, Shen, and Xu 2011) that relates the computation of $Prox_{\frac{\lambda}{L}R_p}$ with the computation of $Prox_{\mu R_2}$ to our circumstance as follows:

Theorem 4. *Let $\boldsymbol{u} \in \mathcal{R}^M$, $c > 0$, and define the mapping $H : \mathcal{R}^N \mapsto \mathcal{R}^N$ as*

$$H(\boldsymbol{v}) = \left(\boldsymbol{I} - Prox_{\frac{\lambda}{cL}R_2}\right)\left(\left(\boldsymbol{I} - c\boldsymbol{B}\boldsymbol{B}^\top\right)\boldsymbol{v} + \boldsymbol{B}\boldsymbol{u}\right). \tag{26}$$

Then, the following equation holds if and only if $\boldsymbol{v} \in \mathcal{R}^N$ is a fixed-point of H.

$$Prox_{\frac{\lambda}{L}R_p}(\boldsymbol{u}) = \boldsymbol{u} - c\boldsymbol{B}^\top\boldsymbol{v}. \tag{27}$$

By the theorem above, computing $Prox_{\frac{\lambda}{L}R_p}$ relies on computing the fixed-point of H. Note that there may be more than one fixed-point of H, but each of them gives the same proximity of $\frac{\lambda}{L}R_p$ at $\boldsymbol{u}$. By simple adaptation, it can be seen from (Micchelli, Shen, and Xu 2011) that H is nonexpansive and it maps $\mathcal{R}^N$ into a closed and convex set $\mathcal{C} \subset \mathcal{R}^N$ that contains $\boldsymbol{0}$, given that R_2 is a Lipschitz continuous convex function and c satisfies the inequality $\left\|\boldsymbol{I} - c\boldsymbol{B}\boldsymbol{B}^\top\right\|_2 \leq 1$.

We recall that for an initial point $\boldsymbol{v}^0 \in \mathcal{R}^N$, the Picard sequence of an operator ψ is defined as $\boldsymbol{v}^{n+1}$=$\psi(\boldsymbol{v}^n)$. If we choose an initial point in $\mathcal{C}$ for H, for example, we set $\boldsymbol{v}^0 = \boldsymbol{0}$, the Picard sequence of H will always stay in $\mathcal{C}$. Therefore, by Opial κ-averaged theorem (Opial 1967), the Picard sequence of κ-averaged operator $H_\kappa = \kappa\boldsymbol{I} + (1-\kappa)H$ converges to a fixed-point of H, $\kappa \in (0, 1)$.

We summarize the algorithm for solving the UGSR model in Algorithm 2, where $\boldsymbol{P}_g$ is the g-th row vector of the distribution matrix $\boldsymbol{P}$ for any $g \in \{1..G\}$, and $diag(\boldsymbol{P}_g)$ indicates a diagonal matrix whose diagonal elements are $\boldsymbol{P}_g^i$ ($i \in \{1..M\}$).

Applications and Experimental Results

We apply UGSR to two classification tasks: text emotion classification and aging face recognition. For any application, in addition to $\boldsymbol{D}$, the distribution matrix $\boldsymbol{P}$ must be given. Actually, the determination of $\boldsymbol{P}$ is domain-dependent. In our two tasks, the setting of $\boldsymbol{P}$ is quite different. For comparison, we also apply SR and GSR to the two tasks, and use the fuzzy k-NN (FkNN) method as a baseline classifier. For the three classifiers under the sparse representation framework, we set the parameter $\lambda = 1$ and optimize them by the proximal method. For FkNN, we set $k = 1$.

Table 1: Texts ("T"+ID from 619 to 624) annotated by six emotions.

EMOTIONS	T619	T620	T621	T622	T623	T624
ANGER	0	26	54	0	0	23
DISGUST	2	9	73	0	2	4
FEAR	12	11	08	3	8	46
JOY	8	2	4	23	13	8
SADNESS	0	31	39	2	4	44
SURPRISE	26	18	17	26	19	14

Table 2: Averaged performance of emotion classification.

	STATISTICS	UGSR	GSR	SR	FkNN
HARD	ACCURACY	**81.57**%	80.11%	80.32%	75.07%
	PRECISION	**39.10**%	33.84%	34.69%	27.33%
	RECALL	**36.63**%	31.55%	31.76%	24.41%
	F1	**37.82**	32.66	33.16	25.79
SOFT	DISTANCE	**0.8426**	0.8797	0.8769	1.1299

We use *Accuracy* (*Acc*), *Precision* (*Pre*), *Recall* (*Rec*) and *F1-measure* (*F1*) to evaluate hard classification performance. *F1* is a combined measure of *Pre* and *Rec*, which is evaluated as $F1$=$\frac{2*Pre*Rec}{Pre+Rec}$. We first calculate these performance measures for each class, and then evaluate their macro-average values across all classes as $Measure_{macro}$=$\frac{1}{G}\sum_{g=1}^{G} Measure_g$.

In soft classification, for each test sample we first estimate the probability distribution via Eq. (19) for UGSR and Eq. (20) for GSR and SR. Then, we calculate the distance (ℓ_1-norm is used as the distance measure) between each method's output and the ground truth distribution, and finally get the averaged result across all test samples.

Text emotion classification

We use the AffectiveText dataset (Strapparava and Mihalcea 2007) for text emotion classification: given some training texts and a set of prespecified emotions, each of these training texts is semantically related to each emotion to a certain degree. For a test text, the hard classification task is to predict the most-likely emotion of the test text, while the soft classification task is to evaluate the probability distribution of the test text semantically related to all emotions.

The dataset contains 1250 short texts, each of which is annotated with the six Eckman emotions: anger, disgust, fear, joy, sadness and surprise, which are marked as *Emotion* 1, ..., *Emotion* 6. Table 1 lists some text samples, where each value means the aggregated frequency of an emotion's keywords appearing in a text. For a training text

Table 3: Averaged performance of aging face recognition.

	c	UGSR	GSR	SR	FkNN	c	UGSR	GSR	SR	FkNN	c	UGSR	GSR	SR	FkNN
ACC	2	**67.50**%	64.00%	61.00%	65.50%	12	**87.68**%	86.32%	86.19%	85.68%	40	**95.70**%	95.39%	95.37%	95.23%
PRE		**55.75**%	50.50%	46.50%	53.75%		**18.13**%	10.70%	9.86%	7.91%		**95.70**%	95.39%	95.37%	95.23%
REC		**67.50**%	64.00%	61.00%	65.50%		**26.08**%	17.92%	17.17%	14.08%		**9.56**%	4.37%	4.14%	2.04%
F1		**61.06**	56.45	52.77	59.05		**21.39**	13.40	12.53	10.13		**11.37**	5.60	5.33	2.82
ACC	6	**78.61**%	76.83%	76.39%	74.50%	20	**92.11**%	91.11%	91.11%	90.92%	82	**97.80**%	97.68%	97.68%	97.62%
PRE		**26.14**%	20.28%	19.05%	14.39%		**14.13**%	6.16%	6.04%	5.04%		**7.72**%	2.87%	3.68%	0.81%
REC		**35.83**%	30.50%	29.17%	23.50%		**21.10**%	11.15%	11.10%	9.25%		**9.76**%	4.88%	4.88%	2.44%
F1		**30.23**	24.36	23.05	17.85		**16.92**	7.94	7.83	6.52		**8.62**	3.61	4.19	1.22

i, $\boldsymbol{P}_g^i$=$\frac{Emotion_g^i}{\sum_{g=1}^{6} Emotion_g^i}$, where $Emotion_g^i$ is the aggregated frequency of *Emotion* g's keywords appearing in text i. For classification using SR and GSR, each training text is assigned to the emotion with the highest probability, this label setting is consistent with the hard classification task that labels each test text with the most-likely emotion.

The dataset has 3082 words in total and we construct 1250 feature vectors of 3082 dimensions using term frequency for all four classifiers. We perform 10-fold cross-validation and compare their outputs with the ground truth.

For hard classification, we compute *Accuracy* and *F1* for 10 times and average the results for each classifier. The averaged results are shown in the first part of Table 2. For soft classification, we compute the distance between each classifier's output and the ground truth distribution using ℓ_1-norm for 10 times, the averaged distances are shown in the second part of Table 2. From Table 2, we can see that UGSR outperforms the other three classifiers, the major reason is that UGSR properly exploits the uncertain group structure information. We also note that SR performs sightly better than GSR, this is because texts in the AffectiveText dataset are not strongly group-sparse: emotions implied in texts are heavily mixed.

Aging face recognition

Here we use the FG-NET dataset [1] for aging face recognition. The task is like this: there are some persons' photos taken at different ages, which are used as training samples. We are now given some of their latest photos, the task is to recognize the corresponding person of each of those photos. This is a hard classification task because we need only to assign one person to each test photo.

The FG-NET dataset contains 1002 different face photos from 82 persons. There are at least 7 face photos for each person at age from 1 to 69 years old. We use the photo of each person taken at her/his oldest age as the test photo and the rest 920 photos as training photos. If we do not consider age information, this task is simply a traditional face recognition task. However, it is a well-known fact that a person's face changes as s/he grows old, and for most people, photos taken at older ages are more similar to the latest photos than the photos taken at younger ages. Considering this, for the g-th person in the training set, let $MaxAge_g$ be the age corresponding to her/his latest photo, we define the probability of the person's photo i taken at age Age_i similar to her/his latest photo as $\boldsymbol{P}_g^i$=$\frac{Age_i}{MaxAge_g}$, and $\boldsymbol{P}_j^i$=0 ($j \neq g$). As each photo can belong to only one person correctly and $\boldsymbol{P}_g^i$ itself does not constitute a distribution, we create an *artificial* label 'OTHER', which indicates anyone but the *correct* person. For any photo i of person g, we have $\boldsymbol{P}_{OTHER}^i$=1-$\boldsymbol{P}_g^i$, which means the probability of photo i not being person g. However, when using the classification rules (3) and (18), we omit the influence of the 'OTHER' label, because 'OTHER' predictions are not considered.

We first resize all photos to bitmaps of 500 pixels to 400 pixels with only gray scales. By extracting Gabor face feature vectors as in (Yang and Zhang 2010) from those resized photos, we get a feature vector of 121520 dimensions for each face photo. Then, we use principal component analysis to reduce the dimensionality of all face vectors to 500.

We evaluate the classifiers by considering different numbers of classes (persons). Given c classes, we repeat the following steps 100 times: first, randomly select c classes from the total 82 classes, then run the four classification methods and record their macro-averaged *Acc*, *Pre*, *Rec* and *F1* values. After the 100 times are over, we average the *Acc*, *Pre*, *Rec* and *F1* values. Here, c is set to 2, 6, 12, 20, 40 and 82, respectively. The results are presented in Table 3. From the experimental results, we can see that 1) as the number of classes increases, all classifiers show worsening performance, which is reasonable. 2) In all settings, UGSR demonstrates advantage over the other three classifiers because it takes advantage of the probability information of all photos taken at different ages. 3) GSR outperforms SR because the training samples are strongly group-aware: training photos of different persons demonstrate considerable discrepancy.

Conclusion

In this paper, we propose a new group sparsity model UGSR to learn group sparse representations from datasets with uncertain group structures, which is a generalization of the standard group sparse representation. An efficient algorithm based on proximity operator is developed to solve the UGSR model. To demonstrate the effectiveness and advantage of the new model, we apply UGSR to two classification tasks: text emotion classification and aging face recognition, where training samples are probabilistically related to different class labels. Experimental results validate the advantage of UGSR over standard GSR and SR as well as fuzzy kNN.

[1] FG-NET database: http://sting.cycollege.ac.cy/ alanitis/fgnetaging/.

Acknowledgments

This work was supported by the Key Projects of Fundamental Research Program of Shanghai Municipal Commission of Science and Technology under grant No. 14JC1400300.

References

Argyriou, A.; Baldassarre, L.; Morales, J.; and Pontil, M. 2011. A general framework for structured sparsity via proximal optimization. *arXiv preprint arXiv:1106.5236*.

Bach, F.; Jenatton, R.; Mairal, J.; and Obozinski, G. 2011. Optimization with sparsity-inducing penalties. *arXiv preprint arXiv:1108.0775*.

Chen, Y.-C.; Patel, V. M.; Pillai, J. K.; Chellappa, R.; and Phillips, P. J. 2013. Dictionary learning from ambiguously labeled data. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 353–360. IEEE.

Hoyer, P. O. 2003. Modeling receptive fields with non-negative sparse coding. *Neurocomputing* 52:547–552.

Huang, J. Z., and Zhang, T. 2009. The benefit of group sparsity. In *Proceedings of the 26th Annual International Converence on Machine Learning*, 417–429.

Jacob, L.; Obozinski, G.; and Vert, J.-P. 2009. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 433–440. ACM.

Jenatton, R.; Audibert, J.-Y.; and Bach, F. 2011. Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research* 12:2777–2824.

Keller, J. M.; Gray, M. R.; and Givens, J. A. 1985. A fuzzy k-nearest neighbor algorithm. *Systems, Man and Cybernetics, IEEE Transactions on* (4):580–585.

Li, Y., and Ngom, A. 2012. Fast sparse representation approaches for the classification of high-dimensional biological data. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, 1–6. IEEE.

Majumdar, A., and Ward, R. 2009. Classification via group sparsity promoting regularization. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 861 –864.

Micchelli, C. A.; Shen, L.; and Xu, Y. 2011. Proximity algorithms for image models: denoising. *Inverse Problems* 27(4):045009.

Moreau, J.-J. 1962. Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A Math* 255:2897–2899.

Olshausen, B. A., and Field, D. J. 1997. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research* 37(23):3311–3325.

Opial, Z. 1967. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society* 73(4):591–597.

Sainath, T. N.; Maskey, S.; Kanevsky, D.; Ramabhadran, B.; Nahamoo, D.; and Hirschberg, J. 2010. Sparse Representations for Text Categorization. In *INTERSPEECH'10*, 2266–2269.

Strapparava, C., and Mihalcea, R. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, 70–74. Association for Computational Linguistics.

Wright, J.; Yang, A.; Ganesh, A.; Sastry, S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31(2):210 –227.

Yang, M., and Zhang, L. 2010. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In *Computer Vision–ECCV 2010*. Springer. 448–461.

Yuan, M., and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49–67.

Yuan, L.; Woodard, A.; Ji, S.; Jiang, Y.; Zhou, Z.-H.; Kumar, S.; and Ye, J. 2012. Learning sparse representations for fruit-fly gene expression pattern image annotation and retrieval. *BMC bioinformatics* 13(1):107.

Zhao, P.; Rocha, G.; and Yu, B. 2006. Grouped and hierarchical model selection through composite absolute penalties. *Department of Statistics, UC Berkeley, Tech. Rep* 703.