$l_{2,1}$ Regularized Correntropy for Robust Feature Selection

Ran He[1], Tieniu Tan[1], Liang Wang[1], Wei-Shi Zheng[2]
[1]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{rhe, tnt,wangliang}@nlpr.ia.ac.cn
[2] School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510275, China
wszheng@ieee.org

Abstract

In this paper, we study the problem of robust feature extraction based on $l_{2,1}$ regularized correntropy in both theoretical and algorithmic manner. In theoretical part, we point out that an $l_{2,1}$-norm minimization can be justified from the viewpoint of half-quadratic (HQ) optimization, which facilitates convergence study and algorithmic development. In particular, a general formulation is accordingly proposed to unify l_1-norm and $l_{2,1}$-norm minimization within a common framework. In algorithmic part, we propose an $l_{2,1}$ regularized correntropy algorithm to extract informative features meanwhile to remove outliers from training data. A new alternate minimization algorithm is also developed to optimize the non-convex correntropy objective. In terms of face recognition, we apply the proposed method to obtain an appearance-based model, called Sparse-Fisherfaces. Extensive experiments show that our method can select robust and sparse features, and outperforms several state-of-the-art subspace methods on large-scale and open face recognition datasets.

In the pattern recognition and computer vision community, feature selection is a fundamental and important method, which aims to select a subset of relevant features meanwhile remove irrelevant and redundant ones out of high-dimensional features. Feature selection can improve generalization capability and speed up learning process [11]. It also helps people better understand about data properties from the curse of dimensionality.

In the past decades, various feature selection methods have been developed [6], among which sparsity regularization is recently considered as one of the most popular ones due to its effectiveness, robustness and efficiency. In l_1-SVM (Support Vector Machine), an l_1-norm regularization is incorporated in SVM to perform feature selection [2]. To form a more structured regularization, Wang et al. [16] propose a hybrid huberized SVM (HHSVM) by combining both l_1-norm and l_2-norm. Since HHSVM is only

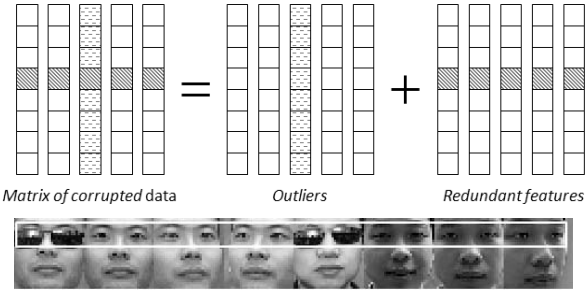


Figure 1. A general framework for robust feature selection. First row: for a corrupted data matrix, we alternately remove outliers and redundant features. Second row: an illustration on the PEAL dataset. If two outliers corrupted by sunglasses are removed from the dataset, features in the white box will be the most discriminative to classify different individuals.

for binary classification, Argyriou et al. [1] further develop a similar $l_{2,1}$ regularized model to deal with feature selection problem in multi-task learning. Recently, Nie et al. [11] propose a robust feature selection method by imposing joint $l_{2,1}$-norm minimization on both loss function and regularization. A new iterative method is also proposed to efficiently optimize the $l_{2,1}$-norm minimization. Based on [11], Gu et al. [5], Hou et al. [9], and Yang et al. [20] apply the joint $l_{2,1}$-norm minimization into subspace learning, sparse regression, and discriminative feature selection respectively. Although different methods based on $l_{2,1}$-norm minimization are proposed, the relationship between the optimal procedure in [11] and other methods (such as iteratively reweighted least squares and half-quadratic optimization) remains unclear. Further theoretical analysis is thus necessary.

Toward this end, this paper presents both theoretical exploration and algorithmic development on $l_{2,1}$-norm minimization. First, a half-quadratic analysis is given for $l_{2,1}$-norm minimizations. Based on this analysis, we can easily extend an $l_{2,1}$-norm loss function to other loss functions and develop new algorithms. Then we present a general frame-

work to unify l_1-norm or $l_{2,1}$-norm regularized robust learning methods. Considering that there are outliers in training data, an $l_{2,1}$ regularized correntropy algorithm is accordingly proposed to extract informative features meanwhile to remove outliers. A new alternate minimization algorithm is also developed to solve the non-convex correntropy objective. Fig. 1 shows the alternate procedure of our method. In each iteration, our method firstly removes outliers and then selects informative features. Finally, we apply the proposed method to face recognition, leading to an appearance-based model, named Sparse-Fisherface (S-Fisherface). Extensive experimental results demonstrate that the proposed method can not only select robust and sparse features, but also perform better than other state-of-the-art subspace methods on large-scale and open face recognition datasets.

Main contributions of this work lie in three-folds:

1) A general framework is proposed for regularized robust learning. It unifies previous l_1 or $l_{2,1}$ regularized robust methods into a general formulation and provides a prelimary platform to develop new methods.

2) An $l_{2,1}$ regularized correntropy model is defined for robust feature selection. Different from the method in [11] that recovers corrupted regression targets, the proposed method removes corrupted samples during learning as shown in Fig. 1.

3) A new appearance-based method (S-Fisherface) is proposed for robust face recognition, which combines feature selection into discriminant subspace learning. As shown in Fig. 3, S-Fisherface learns informative and sparse features against traditional appearance models.

The rest of this paper is organized as follows. We first give a theoretical analysis of $l_{2,1}$ minimization from the view point of HQ, and present a HQ framework for robust feature selection in Section 1. In Section 2, we propose an $l_{2,1}$ regularized correntropy model and examine its application to face recognition. Section 3 provides experimental results, prior to summery of this paper in Section 4.

1. $l_{2,1}$-norm Minimization

This section starts with the study from the half-quadratic analysis for $l_{2,1}$-norm, followed by a general half-quadratic framework for robust feature selection, which unifies l_1- and $l_{2,1}$-norm minimization based robust learning methods. We follow the notations in [11]. Matrices are written as boldface uppercase letters, and vectors are written as boldface lowercase letters. For a matrix $\mathbf{M} = (m_{ij})$, its i-th row is denoted by $\mathbf{m}^i$.

1.1. Half-quadratic Analysis for $l_{2,1}$-norm

In $l_{2,1}$-norm based feature selection methods [11][5], one often aims to solve the following constrained $l_{2,1}$-norm minimization problem,

$$\min_{\mathbf{U}} ||\mathbf{U}||_{2,1} \quad s.t. \quad \mathbf{X}^T\mathbf{U} = \mathbf{Y} \tag{1}$$

where $\|.\|_{2,1}$ is an $l_{2,1}$-norm, projection matrix $\mathbf{U} \in R^{d\times c}$, data matrix $\mathbf{X} \in R^{d\times n}$, and label matrix $Y \in R^{n\times c}$. n is the number of training samples, d is the number of feature dimension, and c is the number of classes. Since the minimizer function of $l_{2,1}$-norm is unpredictable near the origin as shown in Fig. 2 (b), the following objective is often used,

$$\min_{\mathbf{U}} \sum\nolimits_i^d \sqrt{\varepsilon + \|\mathbf{u}^i\|_2^2} \quad s.t. \quad \mathbf{X}^T\mathbf{U} = \mathbf{Y} \tag{2}$$

where ε is a smoothing term. If a decreasing value of ε is used, it can be justified that the algorithm to solve (2) converges to the global solution of (1) [5].

If we define $\phi(x) = \sqrt{\varepsilon + x^2}$, we obtain a general formulation of (2),

$$\min_{\mathbf{U}} \sum\nolimits_i^d \phi(\left\|\mathbf{u}^i\right\|_2) \quad s.t. \quad \mathbf{X}^T\mathbf{U} = \mathbf{Y} \tag{3}$$

In this work, we consider a general case of ϕ that satisfies,

$$\begin{aligned} &x \to \phi(x) \text{ is convex on R,} \\ &x \to \phi(\sqrt{x}) \text{ is concave on } R_+, \\ &\phi(x) = \phi(-x), \forall \mathrm{x} \in \mathrm{R}, \\ &\phi(x) \text{ is } C^1 \text{ on R,} \\ &\phi^{''}(0^+) > 0, \ \lim_{x\to\infty} \phi(x)/x^2 = 0. \end{aligned} \tag{4}$$

It is easy to prove that $\phi(x) = \sqrt{\varepsilon + x^2}$ satisfies all conditions in (4). The following Lemma 1 founds the base for optimizing $\phi(.)$ in a half quadratic way [12].

Lemma 1. *Let $\phi(.)$ be a function satisfying all conditions in (4), there exists a conjugate function $\varphi(.)$ (or named dual potential function in [12]), such that*

$$\phi(\left\|\mathbf{u}^i\right\|_2) = \inf_{p\in R} \left\{ p\left\|\mathbf{u}^i\right\|_2^2 + \varphi(p) \right\} \tag{5}$$

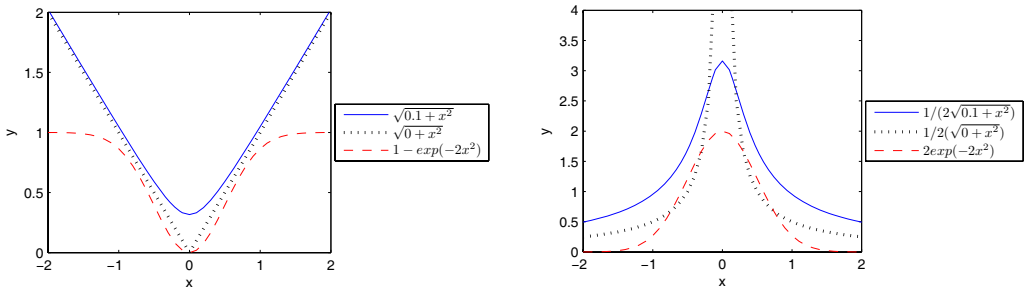
where p is determined by the minimizer function $\delta(.)$ with respect to $\phi(.)$.

To solve the general constrained optimization problem in (3), we firstly introduce Lagrange multipliers Λ, giving the following Lagrangian function,

$$\mathcal{L}(\mathbf{U}) = \sum\nolimits_{i=1}^d \phi(\left\|\mathbf{u}^i\right\|_2) - Tr(\Lambda^T(\mathbf{X}^T\mathbf{U} - \mathbf{Y})) \tag{6}$$

where $Tr(.)$ is the matrix trace operator. Using (5) on each $\phi(\left\|\mathbf{u}^i\right\|_2)$ for i, the augmented cost-function $\mathcal{J}$ of (6) reads,

$$\mathcal{J}(\mathbf{U}, \mathbf{q}) = Tr(\mathbf{U}^T\mathbf{Q}\mathbf{U}) - Tr(\Lambda^T(\mathbf{X}^T\mathbf{U} - \mathbf{Y})) \tag{7}$$



(a) loss function (b) minimizer function

Figure 2. Potential loss functions and their corresponding minimizer functions in Half-quadratic optimization.

where $\mathbf{q} \in R^d$ is an auxiliary vector, and $\mathbf{Q} = diag(\mathbf{q})$. The operator $diag(.)$ puts a vector $\mathbf{q}$ on the main diagonal of $\mathbf{Q}$. According to Lemma 1, it can be drawn immediately that for a fixed $\mathbf{U}$, $\mathcal{L}(\mathbf{U}) = \min_{\mathbf{q}} \mathcal{J}(\mathbf{U}, \mathbf{q})$. Based on the Half-quadratic optimization, $\mathcal{J}(\mathbf{U}, \mathbf{q})$ can be solved by the following alternate minimization way,

$$\mathbf{q}_i^t = \delta(\|\mathbf{u}^i\|_2) \quad (8)$$

$$\mathbf{U}^t = \arg\min_{\mathbf{U}} \mathcal{J}(\mathbf{U}, \mathbf{q}^t) \quad (9)$$

where $\delta(.)$ is the minimizer function with respect to $\phi(.)$.

Setting the derivative of $\mathcal{J}(\mathbf{U}, \mathbf{q})$ with respect to $\mathbf{U}$ to zero, we obtain

$$\frac{\partial \mathcal{J}(\mathbf{U}, \mathbf{q})}{\partial \mathbf{U}} = 2\mathbf{Q}\mathbf{U} - \mathbf{X}\Lambda = 0 \quad (10)$$

Left multiplying both sides of (10) by $\mathbf{X}^T\mathbf{Q}^{-1}$, and using the equality constraint $\mathbf{X}^T\mathbf{U} = Y$, we have:

$$\begin{aligned} &2\mathbf{X}^T\mathbf{U} - \mathbf{X}^T\mathbf{Q}^{-1}\mathbf{X}\Lambda = 0 \\ &\Rightarrow 2\mathbf{Y} - \mathbf{X}^T\mathbf{Q}^{-1}\mathbf{X}\Lambda = 0 \\ &\Rightarrow \Lambda = 2(\mathbf{X}^T\mathbf{Q}^{-1}\mathbf{X})^{-1}\mathbf{Y} \end{aligned} \quad (11)$$

Then we obtain the analytic solution of (9):

$$\mathbf{U}^* = \mathbf{Q}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{Q}^{-1}\mathbf{X})^{-1}\mathbf{Y} \quad (12)$$

Algorithm 1: Feature Selection via $l_{2,1}$-norm

Input: $\mathbf{X} \in R^{d\times n}$ and $\mathbf{Y} \in R^{n\times c}$.
Output: $\mathbf{U} \in R^{d\times c}$

1: $\mathbf{U} \leftarrow 0$ and $t \leftarrow 1$.
2: **repeat**
3: Compute the auxiliary vector $\mathbf{p}^t$ according to (8).
4: Compute $\mathbf{U}^t$ according to (12).
5: **until** Converges

We summarize our proposed alternate minimization algorithm to optimize (3) in Algorithm 1. The HQ formulation facilities the convergence proof of $l_{2,1}$ minimization procedure. Proposition 1 confirms that the sequences of Algorithm 1 will converge. Since the proof is trivial, we omit it from this paper due to page limit. The experimental analysis in [12] shows that HQ based methods potentially run faster than quasi-Newton and steepest descent schemes.

Proposition 1. *Denote* $\mathcal{J}^t = \mathcal{J}(\mathbf{U}^t, \mathbf{q}^t)$, *the sequences* $\{\mathcal{J}^t, t = 1, 2, \dots\}$, $\{\mathbf{U}^t, t = 1, 2, \dots\}$ *and* $\{\mathbf{q}^t, t = 1, 2, \dots\}$ *generated by (8) and (12) converge.*

Fig. 2 (a) shows three potential functions of $\phi(.)$ that are used in compressed sensing as an approximation of l_0-norm. Fig. 2 (b) shows their corresponding minimizer functions. In robust regression and iteratively reweighted least squares (IRLS) [22][10], minimizer function is often called weighting function. Note that the absolute function $\phi_1(x) = |x| = \sqrt{x^2}$ in l_1-norm has not a minimizer function in HQ optimization but only has a weighting function in IRLS. Since its weighting function is unpredictable near the origin, $\phi_2(x) = \sqrt{\varepsilon + x^2}$ is often used as an approximation of $|x|$ [5]. Function $\phi_3(x) = 1 - exp(-2x^2)$ is used in correntropy as an approximation of l_0-norm [14].

Comparing curves of three potential functions, we see that within the range $[-1, 1]$, functions $\phi_2(.)$ and $\phi_3(.)$ can be treated as an approximation of $\phi_1(.)$. But the minimizer functions of $\phi_2(.)$ and $\phi_3(.)$ are predicable near the origin. When $|x| > 1$, things become different. Function $\phi_3(.)$ gives the same loss value (i.e., $\phi_3(x) = 1$) whereas the other two functions do not. This character makes the minimizer function of $\phi_3(.)$ in Fig.2 (b) has value 0 when $|x| > 1.5$, which may be helpful in optimization to solve real-world problems.

1.2. A General Half-quadratic Framework for Robust Feature Selection

Robust learning has drawn much attention in machine learning and computer vision [17]. Many methods have been developed to deal with outliers in training or testing sets. Considering the HQ analysis for $l_{2,1}$-norm mentioned

Function	Variable	Objective	Method
$\phi_o = \phi_R = x^2$	$\mathbf{u} \in R^d, \mathbf{y} \in R^n$	$\min_{\mathbf{u}} \left\|\mathbf{X^T u} + \mathbf{y}\right\|_2^2 + \lambda \left\|\mathbf{u}\right\|_2^2$	Rigid regression [3]
$\phi_o = x^2, \phi_R = \|x\|$	$\mathbf{u} \in R^n, \mathbf{y} \in R^d$	$\min_{\mathbf{u}} \left\|\mathbf{X u} + \mathbf{y}\right\|_2^2 + \lambda \left\|\mathbf{u}\right\|_1$	LASSO [15][17]
$\phi_o = \phi(x), \phi_R = x^2$	$\mathbf{u} \in R^d, \mathbf{y} \in R^n$	$\min_{\mathbf{u}} \sum_i \phi_o((\mathbf{X^T u} + \mathbf{y})^i) + \lambda \left\|\mathbf{u}\right\|_2^2$	Robust regression [22][10][21]
$\phi_o = \phi(x), \phi_R = \|x\|$	$\mathbf{u} \in R^n, \mathbf{y} \in R^d$	$\min_{\mathbf{u}} \sum_i \phi_o((\mathbf{X u} + \mathbf{y})^i) + \lambda \left\|\mathbf{u}\right\|_1$	Robust sparse representation [7][8][19]
$\phi_o = \phi_R = \sqrt{x^2}$	$\mathbf{U} \in R^{d \times m}$	$\min_{\mathbf{U}} \left\|\mathbf{X^T U} + \mathbf{Y}\right\|_{2,1} + \lambda \left\|\mathbf{U}\right\|_{2,1}$	Robust feature selection [11][9][20][5]

Table 1. Summary of some special cases of our proposed framework. These cases have been widely used in computer vision and machine learning. $\phi(.)$ is the potential function that satisfies (4) .

above, we regard a general robust learning problem, i.e.,

$$\min_{\mathbf{U}} \sum_{i=1}^{d} \phi_o(\left\|(\mathbf{AU} + \mathbf{B})^i\right\|_2) + \lambda \sum_{i=1}^{d} \phi_R(\left\|\mathbf{u}^i\right\|_2) \quad (13)$$

where $\phi_o(.)$ and $\phi_R(.)$ satisfy all of conditions in (4). According to HQ minimization, we can solve (13) by the following alternate minimization way,

$$\mathbf{p}_i^t = \delta_o(\left\|(\mathbf{AU} + \mathbf{B})^i\right\|_2) \quad (14)$$

$$\min_{\mathbf{U}} \left\{ \sum_{i=1}^{d} \mathbf{p}_i^t \left\|(\mathbf{AU} + \mathbf{B})^i\right\|_2^2 + \lambda \sum_{i=1}^{d} \phi_R(\left\|\mathbf{u}^i\right\|_2) \right\} (15)$$

where $\delta_o(.)$ is the minimizer function of $\phi_o(.)$ in HQ optimization. According to HQ optimization, the above iterative procedure monotonously decreases until it converges.

Table 1 lists some special cases of (13), which are commonly used in the recent literature. We firstly consider the matrix variable $\mathbf{U}$ as a vector variable $\mathbf{u}$. When $\phi_o(x) = x^2$, the objective in (13) becomes the standard regularized least squares problem. When $\phi_R(.) = x^2$ and $\phi_R(.) = |x|$, the objectives become rigid regression and LASSO respectively. The former is widely used in spectral regression [3]; and the later is widely used in sparse representation [17].

The third and fourth rows of Table 1 show the objectives used for robust regression and robust sparse representation respectively, where $\phi(.)$ satisfies (4) and also belongs to M-estimators. Both of the objectives are solved via the iterative procedure in (14) and (15). In robust regression [22][10][21], ones firstly calculate weights according to (14), and then solve a weighted least squares problem in (15). In robust sparse representation, ones firstly calculate weights (or select informative features) according to (14), and then solve a weighted l_1 minimization problem in (15).

When $\mathbf{U}$ is a matrix and $\phi_o = \phi_R = \sqrt{x^2}$, the objective in (13) becomes the objective used in robust feature selection, as shown in the fifth row of Table 1. Our half-quadratic analysis for $l_{2,1}$ minimization in Section 1.1 can be viewed as a complement of [11]. From the HQ viewpoint, robust feature selection in [11] can be viewed as an extension of robust sparse representation in [7][8][19] from vector variable to matrix variable. Both of these two categories of methods harness the minimizer function of HQ to select informative features, and then perform learning.

2. $l_{2,1}$ Regularized Correntropy for Robust Feature Selection

In this section, we firstly propose an $l_{2,1}$ regularized correntropy model for robust feature selection. Different from the method in [11] that estimates the errors in label matrix target Y, our method iteratively removes outliers in the training set by reweighting. Then we apply the proposed model in appearance based face recognition.

2.1. $l_{2,1}$ Regularized Correntropy

Correntropy is proposed in information theoretic learning to process non-Gaussian noise and impulsive noise [10], and is widely used in computer vision and signal processing. It is directly related to the Renyi's quadratic entropy, and has a close relationship with Welsch M-estimators. Considering that correntropy tends to control outliers better than other M-estimators [10][7], we develop an $l_{2,1}$ regularized correntropy model for robust feature selection. Fig. 1 shows our basic motivation. By applying correntropy and $l_{2,1}$ regularization in (13), we obtain the following correntropy objective,

$$\min_{\mathbf{U}} \{1 - \sum_{k=1}^{n} \exp(-\frac{\left\|(\mathbf{X}^T\mathbf{U} - \mathbf{Y})^k\right\|_2^2}{\sigma^2}) + \left\|\mathbf{U}\right\|_{2,1}\} \quad (16)$$

where σ is the kernel size that controls all properties of correntropy. In the objective, the correntropy is used to remove outliers and the $l_{2,1}$ regularization is used to select robust and informative features.

According to the HQ in Section 1.2, the above objective can be solved in an alternate minimization way as follows,

$$p_k^t = \exp(-\left\|(\mathbf{X}^T\mathbf{U} - \mathbf{Y})^k\right\|_2^2 / \sigma^2) \quad (17)$$

$$q_i^t = 1/(2\left\|\mathbf{u}^i\right\|_2) \quad (18)$$

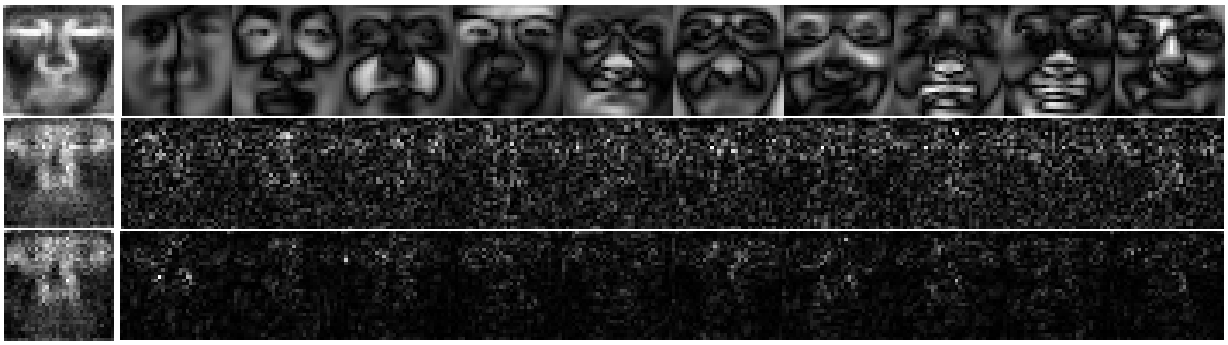


Figure 3. Energy faces (the left most column), Eigenfaces (the first row), Fisherfaces (the second row), and S-Fisherfaces (the third row) calculated from face images on the FRGC database. Since there are positive and negative values in Eigenfaces, Fisherfaces and S-Fisherfaces, we show their absolute values here. The energy faces is computed by (21). The first row: absolute Eigenfaces (PCA). In energy face of Eigenfaces, the ratio of maximum absolute value and minimum one is 2.31. The second row: absolute Fisherfaces (LDA). In energy face of Fisherfaces, the ratio of maximum absolute value and minimum one is 2.99. The third row: absolute S-Fisherfaces ($l_{2,1}$ regularized correntropy). In energy face of S-Fisherfaces, the ratio of maximum absolute value and minimum one is 13.57.

$$\begin{aligned} U^t &= \arg\min_{\mathbf{U}} Tr((\mathbf{X}^T\mathbf{U}-\mathbf{Y})^T\mathbf{P}(\mathbf{X}^T\mathbf{U}-\mathbf{Y})) \\ &+ \lambda Tr(\mathbf{U}^T\mathbf{Q}\mathbf{U}) \end{aligned} \tag{19}$$

Where $\mathbf{p}$ and $\mathbf{q}$ are auxiliary variables of correntropy and $l_{2,1}$-norm respectively. And $\mathbf{P} = diag(\mathbf{p})$ and $\mathbf{Q} = diag(\mathbf{q})$. The analytic solution of (19) is given by,

$$\mathbf{U}^* = (\mathbf{X}\mathbf{P}\mathbf{X}^T + \lambda\mathbf{Q})^{-1}\mathbf{X}\mathbf{P}\mathbf{Y} \tag{20}$$

To save computational cost, the optimal solution of (20) can be computed via solving the linear system problem $(\mathbf{X}\mathbf{P}\mathbf{X}^T + \lambda\mathbf{Q})\mathbf{U} = \mathbf{X}\mathbf{P}\mathbf{Y}$.

Algorithm 2: Correntropy Induced Robust Feature Selection (CRFS)

Input: $\mathbf{X} \in R^{d\times n}$ and $\mathbf{Y} \in R^{n\times c}$.
Output: $\mathbf{U} \in R^{d\times c}$

1: $\mathbf{U} \leftarrow 0$ and $t \leftarrow 0$.
2: **repeat**
3: Compute $p_k^t = \exp(-\left\|(\mathbf{X}^T\mathbf{U}-\mathbf{Y})^k\right\|_2^2/\sigma^2)$
4: Compute $q_i^t = 1/(2\left\|\mathbf{u}^i\right\|_2)$
5: Compute $\mathbf{U}^t$ by solving the linear system: $(\mathbf{X}\mathbf{P}\mathbf{X}^T + \lambda\mathbf{Q})\mathbf{U} = \mathbf{X}\mathbf{P}\mathbf{Y}$.
6: **until** Converges

Algorithm 2 summarizes the alternate minimization procedure to optimize (16). In step 3, we compute the auxiliary vector $\mathbf{p}^t$. If there are outliers in the training set, they will receive small values in $\mathbf{p}^t$ due to the robustness of correntropy. In step 4, we compute the auxiliary vector $\mathbf{q}^t$ that corresponds to $l_{2,1}$-norm and plays a role in feature selection. In step 5, we find the optimal solution $\mathbf{U}^*$. According to HQ optimization, the objective function is minimized in each step. The correntropy objective is bounded, and hence Algorithm 2 will decrease (16) step by step until it converges.

2.2. S-Fisherface (Sparse Fisherface)

In this subsection, we apply $l_{2,1}$ regularized correntropy algorithm to appearance based face recognition. A face image can be mapped into the learned subspace $\mathbf{U}$ and then is classified. We can display the projection vectors in $\mathbf{U}$ as images. Considering that linear discriminant analysis can be treated as a multi linear regression [3][18], we call these images as sparse Fisherfaces (S-Fisherfaces). Using the FRGC face database as the training set, we show the absolute value of S-Fisherfaces in Fig. 3, together with Eigenfaces and Fisherfaces.

It is interesting to see that S-Fisherfaces are somehow similar to Fisherfaces. This may be due to the fact that they are all related to linear regression. However, the intensity of most areas in S-Fisherfaces is darker. A darker value indicates a smaller value. Hence S-Fisherfaces are sparser than Fisherfaces and Eigenfaces. To further analyze different appearance methods, we introduce the concept of energy face.

Give a set of Eigenfaces (or Fisherfaces) $\mathbf{U} \in R^{d\times m}$ in which each column is a Eigenface, the energy face is defined as a vector $\mathbf{e}$ whose item is computed as follows,

$$\mathbf{e}_i = \sum_{j=1}^{m} \mathbf{U}_{ij}^2 \tag{21}$$

Fig. 3 shows these energy faces of the three appearance methods. We see that in S-Fisherfaces more features are selected around eyes and nose. We consider this phenomenon as a coincidence with that the features in face recognition around two eyes and nose are often discriminative.

In energy faces, the ratio of maximum absolute value and minimum one for Eigenface, Fisherface, S-Fisherface are

2.31, 2.99, and 13.57 respectively. A larger ratio indicates a sparser solution. We see that Eigenface and Fisherface have similar sparsity whereas S-Fisherface is sparser. This is due to the $l_{2,1}$-norm regularization in the correntropy objective.

In the energy face of S-Fisherfaces, we observe that most features are around the eyes and nose. The features around mouth are significantly affected by expression variation such that they are less discriminative. Since S-Fisherfaces computed by Algorithm 2 involve robust feature selection during learning, they only contain most informative and discriminative features. Irrelevant and redundant facial features are removed during the alternate minimization. As a result, S-Fisherfaces are potentially better than previous appearance-based methods,

3. Experiments

In this section, several experiments on a couple of large-scale face recognition datasets are carried out to show that our proposed CRFS method (Algorithm 2) has more discriminating power than previous appearance based methods and be less sensitive to outliers. Since real-world face recognition is an open set problem, we make use of training set, probe set, and gallery set to evaluate different methods [13]. Three appearance methods (principal component analysis (PCA), linear discriminant analysis (LDA), and locality preserving projections (LPP) [1]) and two robust methods(Renyis entropy discriminant analysis (REDA) [21] and robust feature selection (RFS) [11]) are compared. The nearest neighbor algorithm based on the Euclidean distance is used as classifier [3].

3.1. Results on the FRGC Database

In this subsection, we evaluate different methods on the large-scale and challenging FRGC database. We collect facial images from a subset of the most challenging FRGC version 2 face database [13]. There are 8014 images of 466 subjects in the query set for the FRGC experiment 4. These uncontrolled images contain variations of illumination, expression, time, and blurring. We take the first 20 facial images if the number of facial images is not less than 20. Then we obtain 3720 facial images of 186 subjects. Each facial image is in 256 gray scales per pixel and cropped into size of 64×64 pixels by fixing the positions of two eyes.

In this first experiment, the first 60 subjects are used as the training set, and the remaining 126 subjects are exploited as the gallery set and the probe set. Then we take the first 10 facial images of each person in the last 126 subjects as the gallery set and the remaining 10 images as the probe set. In the second experiment, the first 120 subjects are used as the training set, and the remaining 66 subjects are exploited as the gallery set and the probe set. Therefore, the set of persons for training is different from that for testing.

[1] http://www.zjucadcg.cn/dengcai/Data/data.html

Number	PCA	LPP	LDA	REDA	RFS	CRFS
60	26.9	21.9	12.5	18.8	11.3	**11.2**
120	26.2	16.8	7.4	13.3	7.1	**6.7**

Table 2. Recognition error rates with different number of training subjects on FRGC dataset. The numbers 60 and 120 indicate that 60 and 120 subjects are used in training set respectively.

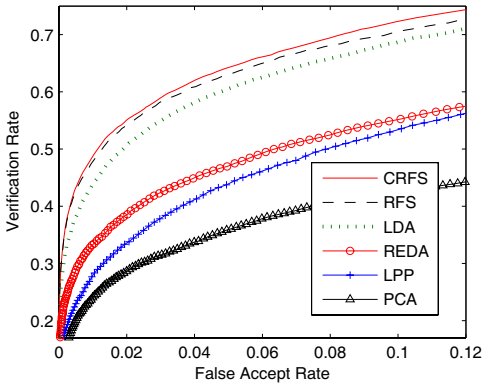


Figure 4. ROC curves of different methods on the FRGC dataset.

Table 2 shows recognition error rates of different methods in these two experiments. We see that the methods can be ordered in descending error rates as PCA, LPP, REDA, LDA, RFS, and CRFS. RFS and CRFS perform better than the remaining three methods. And CRFS slightly outperforms RFS. This may be due to that both of them are based on similar objectives. But CRFS can greatly reduce computational cost, compared with RFS as shown in Section 3.3. We also observe that LPP and REDA perform worse than LDA. This may be because that the set of persons for training is different from that for testing.

Fig. 4 further shows the receiver operator characteristic (ROC) curves of different methods. As expected, CRFS achieves the highest ROC curve. CRFS slightly outperforms RFS. When there are no outliers that are significantly different from other samples, CRFS and RFS seem to achieve similar recognition accuracy.

3.2. Results on the PEAL Database

In this experiment, we evaluate the robustness of different methods on the challenging CAS-PEAL database [4]. The CAS-PEAL database is a large-scale Chinese face database, which contains 99,594 images of 1040 individuals with varying Pose, Expression, Accessory, and Lighting (PEAL). We select all frontal facial images under expression and lighting variations, where all frontal facial images

Scenario	LPP	REDA	LDA	RFS	CRFS
original dataset	56.1	44.0	28.8	23.7	21.4
sunglasses (10%)	56.4±0.3	44.3±0.3	31.4±0.2	24.0±0.1	21.9± 0.1
hat (20%)	58.3±0.2	44.5±0.5	32.5±0.3	24.6±0.1	22.2±0.1
mislabeling (20%)	68.0±1.0	47.6±0.9	39.7±0.5	25.9±0.6	23.5±0.5

Table 3. Recognition error rates under different types of outliers on the PEAL datasets.

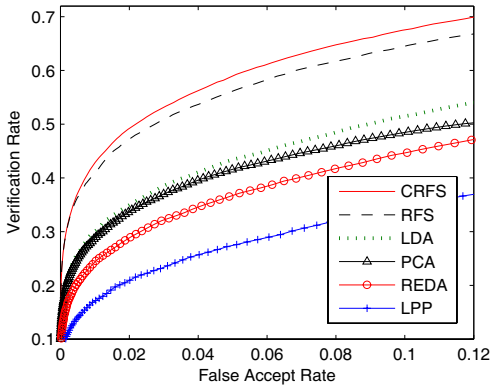


Figure 5. ROC curves of different methods on the PEAL dataset.

whose pose degrees are less than or equal to 22^o. Then the whole dataset contains 7448 images of 1038 individuals. All the selected images are cropped with dimension 32×32, and are enhanced by using histogram equalization.

In the first experiment, for the training set, we select 2226 images of 261 individuals that correspond to the individuals with sunglasses. We take the half facial images of each person in the remaining 777 individuals as the gallery set, the remaining images as the probe set. Therefore, the set of persons for training is different from that for testing. In the second experiment, we introduce three types of outliers. (1) For outliers, we randomly selected 223 images from the sunglasses occluded images of the 261 individuals. The level of noise is 10%. (2) We randomly select 445 images from the hat occluded images of the 261 individuals. The level of noise is 20%. (3) We randomly mislabel 20% labels of the images in the training set. All experiments are repeated 20 times, and mean and deviation are reported.

Table 3 lists the recognition error rates of different methods under different types of outliers on the PEAL dataset. The first row of Table 3 gives recognition rates on the training set without outliers. CRFS significantly outperforms its four competitors. The improvements of CRFS against RFS and LDA are 9.7% and 25.7% respectively. The error rates of LPP and REDA are very high. This may be because the number of the images per person is different in the training set. The local structure of LPP may be inaccurate due to noise such that LPP fails to reflect the true structure on the probe and gallery set. Fig. 5 further shows the ROC curves of different methods on the uncorrupted training set. Both RFS and CRFS significantly perform better than other methods. Due to the same reason discussed above, the ROC curves of LPP and REDA are even lower than that of PCA. But we consider this phenomenon as an agreement with this PEAL data set.

For hat and sunglasses occlusions, error rates of all methods increase. We observe that the increment of the error rate of LDA is larger than those of other methods. This is because LDA calculates intra-class and inter-class matrices during training. Hat or sunglasses occlusions significantly change face appearance and hence they affect the computation of the two matrices. As a result, the error rates of LDA increase larger than those of other methods.

When there is mislabeling noise, the performance of all methods decrease significantly. In particular, the error rates of two non-robust methods LPP and LDA increase rapidly. Since discriminative LPP is based on a local structure which depends on label information, it is sensitive to mislabeling noise. We also observe that the error rates of three robust methods (REDA, RFS, and CRFS) increase slowly. This is due to that they are based on robust M-estimators and detect outliers in each iteration. As expected, CRFS obtains the lowest recognition rates among all compared methods.

3.3. Computational Cost

In many vision problems, there are often tremendous classes in training set. For example, in face recognition, the number of classes is equal to that of persons. When the number of classes tends to be larger, label matrix $\mathbf{Y}$ will be very large. Since RFS iteratively estimates an error matrix $\mathbf{E}$ that has the same size as $\mathbf{Y}$, the computational cost of RFS will tend to be large. Different from RFS, our CRFS makes use of a weighting strategy to deal with outliers. In each iteration, it only computes auxiliary vector $\mathbf{p} \in R^n$. Hence it can speed up learning procedure.

Table 4 tabulates computational costs of different methods. When the number of subjects in the training set is 120, the methods can ordered in descending computational costs as RFS, CRFS, LPP, LDA, REDA, and PCA. We observe that CRFS can significantly reduce the computational cost as compared with RFS, especially when the number of subjects is large.

Number	PCA	LPP	REDA	LDA	RFS	CRFS
60	4.8	6.9	5.9	5.7	6.8	5.8
120	5.5	8.4	6.8	7.2	33.3	10.4

Table 4. Computation time (s) on the FRGC dataset.

4. Conclusion

This paper has studied $l_{2,1}$-norm minimization from the viewpoint of HQ optimization, and proposes a general formulation to unify l_1- and $l_{2,1}$- regularized robust learning methods. Based on the HQ analysis, an $l_{2,1}$ regularized correntropy algorithm has been further presented to extract informative features meanwhile to remove outliers from the training set. An alternate minimization algorithm has been used to optimize the non-convex correntropy objective. Applying the proposed method to face recognition problem, we have obtained a new appearance based face recognition model - Sparse-Fisherface. Extensive experiments have validated that our method can select robust and sparse features, and outperforms other appearance-based methods on large-scale and open face recognition datasets.

5. Acknowledgment

This work is funded by the Research Foundation for the Doctoral Program of the Ministry of Education of China (Grant No. 20100041120009), the Grant of Hundred Talents Program of CAS, the National Basic Research Program of China (Grant No. 2012CB316300), National Natural Science Foundation of China (Grant No. 60736018, 61075024, 61103155, 61175003), International S&T Cooperation Program of China (Grant No.2010DFB14110) and National Key Technology R&D Program (Grant No.2012BAK02B01).

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Neural Information Processing Systems*, pages 41–48, 2007.

[2] P. Bradley and O. Mangasarian. Feature selection via concave minimization and support vector machines. In *International Conference on Machine Learning*, 1998.

[3] D. Cai, X. He, and J. Han. Spectral regression for efficient regularized subspace learning. In *International Conference on Computer Vision*, pages 1–7, 2007.

[4] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao. The cas-peal large-scale Chinese face database and baseline evaluations. *IEEE Transactions on System, Man, and Cybernetics (Part A)*, 38(1):149–161, 2008.

[5] Q. Gu, Z. Li, and J. Han. Joint feature selection and subspace learning. In *International Joint Conferences on Artificial Intelligence*, 2011.

[6] I. Guyon and A. Elissee. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[7] R. He, W.-S. Zheng, and B.-G. Hu. Maximum correntropy criterion for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1561–1576, 2011.

[8] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong. A regularized correntropy framework for robust pattern recognition. *Neural Computation*, 23(8):2074–2100, 2011.

[9] C. Hou, F. Nie, D. Yi, and Y. Wu. Feature selection via joint embedding learning and sparse regression. In *International Joint Conferences on Artificial Intelligence*, pages 1324–1329, 2011.

[10] W. Liu, P. P. Pokharel, and J. C. Principe. Correntropy: Properties and applications in non-Gaussian signal processing. *IEEE Transactions on Signal Processing*, 55(11):5286–5298, 2007.

[11] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint $l_{2,1}$-norms minimization. In *Neural Information Processing Systems*, pages 1813–1821, 2010.

[12] M. Nikolova and M. K. NG. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal on Scientific computing*, 27(3):937–966, 2005.

[13] P. J. Phillips, P. J. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *conference on Computer Vision and Pattern Recognition*, 2005.

[14] S. Seth and J. C. Principe. Compressed signal reconstruction using the correntropy induced metric. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 3845–3848, 2008.

[15] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996.

[16] L. Wang, J. Zhu, and H. Zou. Hybrid huberized support vector machines for microarray classification. In *International Conference on Machine Learning*, 2007.

[17] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of IEEE*, 98(6):1031–1044, 2010.

[18] S. Xiang, F. Nie, and C. Zhang. Semi-supervised classification via local spline regression. *IEEE Transactions Pattern Analysis Machine Intelligence*, 32(11):2039–2053, 2010.

[19] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition. In *conference on Computer Vision and Pattern Recognition*, pages 625–632, 2011.

[20] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou. $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *International Joint Conferences on Artificial Intelligence*, pages 1589–1594, 2011.

[21] X.-T. Yuan and B.-G. Hu. Robust feature extraction via information theoretic learning. In *International Conference on Machine Learning*, pages 1193–1200, 2009.

[22] Z. Zhang. Parameter estimation techniques: A tutorial with application to conic fitting. *Image and Vision Computing*, 15(1):59–76, 1997.