Comput Optim Appl (2014) 58:409–421
DOI 10.1007/s10589-014-9648-x

A unified algorithm for mixed $l_{2,p}$-minimizations and its application in feature selection

Liping Wang · Songcan Chen · Yuanping Wang

Received: 27 June 2013 / Accepted: 20 February 2014 / Published online: 7 March 2014

Abstract Recently, matrix norm $l_{2,1}$ has been widely applied to feature selection in many areas such as computer vision, pattern recognition, biological study and etc. As an extension of l_1 norm, $l_{2,1}$ matrix norm is often used to find jointly sparse solution. Actually, computational studies have showed that the solution of l_p-minimization ($0 < p < 1$) is sparser than that of l_1-minimization. The generalized $l_{2,p}$-minimization ($p \in (0, 1]$) is naturally expected to have better sparsity than $l_{2,1}$-minimization. This paper presents a type of models based on $l_{2,p}$ ($p \in (0, 1]$) matrix norm which is non-convex and non-Lipschitz continuous optimization problem when p is fractional ($0 < p < 1$). For all p in $(0, 1]$, a unified algorithm is proposed to solve the $l_{2,p}$-minimization and the convergence is also uniformly demonstrated. In the practical implementation of algorithm, a gradient projection technique is utilized to reduce the computational cost. Typically different $l_{2,p}$ ($p \in (0, 1]$) are applied to select features in computational biology.

Keywords Mixed matrix norm · Non-Lipschitz continuous · Unified algorithm · Gradient projection

The work is supported by the NSFC11001128, NSFC61035003, NSFC61170151, NSFC11071117 and the Fundamental Research Funds for the Central Universities (No. NZ2013306 and NZ2013211).

L. Wang · Y. Wang
Department of Mathematics, Nanjing University of Aeronautics and Astronautics, Nanjing, China

S. Chen (✉)
Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China
e-mail: s.chen@nuaa.edu.cn

1 Introduction

In many applications, such as computer vision, handwriting character recognition, medical diagnosis and etc., $l_{2,1}$ matrix norm has received increasing attention to select features for its joint sparsity. The underlying assumption is that features in high dimensional data are related to each other but only a few of informational features contribute to discrimination. Existing schemes employ various strategies to capture the feature relatedness and select the most discriminative features, then take into account in the learning formulation. The involved models and methods are frequently constructed in l_1-norm framework. Actually, extensive computational studies [2–5,12] have showed that using l_p-norm ($0 < p < 1$) can find sparser solution than using l_1-norm. Naturally, one can expect mixed $l_{2,p}$-norm ($0 < p \leq 1$) based minimization to have better sparsity pattern than $l_{2,1}$-norm. A similar $l_p - l_q$ ($0 < p \leq 1,\ 1 \leq q \leq 2$) penalty for sparse linear and multiple kernel multi-task learning has been considered in [13]. But the induced optimization problems have to be separately solved by different algorithms according to the convex ($p = 1$) and non-convex ($0 < p < 1$) cases. This brings computational difficulty to freely vary p and q. This paper presents a generalized model based on $l_{2,p}$ ($p \in (0, 1]$) matrix norm[1]. When p is a positive fraction ($0 < p < 1$), the involved optimization problem is neither convex nor Lipschitz continuous. When $p = 1$, it is just the well defined $l_{2,1}$-minimization. Inspired by the work in [1,2], we will develop a unified algorithm to solve the mixed $l_{2,p}$-norm based minimization for all p in $(0, 1]$. To the best of our knowledge, this presentation has the innovations as follows. (1) The general model based on mixed $l_{2,p}$ ($p \in (0, 1]$) norm is more adaptive than $l_{2,1}-$norm to offer better sparsity for different data structures. (2) The unified algorithm and its unform convergence for $p \in (0, 1]$ provides algorithmic support in pursuing more sparse patterns. (3) In the implementation of unified algorithm, a gradient projection scheme is utilized to reduce the computational cost. The running CPU time comparisons confirm the numerical economy of this technique. (4) Typical $p \in (0, 1]$ are tested in $l_{2,p}$-minimization, the experiments in bioinformatics study provide empirical evidence that some $0 < p < 1$ are alternatives in constructing better sparse patterns than $p = 1$.

The rest of the paper is organized as follows. Section 2 states some necessary notations and induces the generalized model based on $l_{2,p}$-norm ($p \in (0, 1]$). Section 3 develops a unified approach to solve the specially mixed optimization problem, and the convergence analysis is also established. Section 4 considers a gradient projection technique for solving the subproblems inexactly. Some experiment results are reported in the Sect. 5. Conclusions and further extensions are discussed in the last section.

2 $l_{2,p}$-Norm based minimization

We employ the notations as usual. Matrices are written as uppercase letters while vectors are written as lowercase letters. For example, $A = (a_{ij})_{d\times n}$ denotes a real

[1] $\|\cdot\|_{2,p}$ ($0 < p < 1$) is not a valid matrix norm because it does not admit the triangular inequality. Here we call it matrix norm for convenience.

$d \times n$ matrix, $a^i \in R^n (i = 1, \cdots, d)$ and $a_j \in R^d (j = 1, \cdots, n)$ are the $i-$th row and $j-$th column of A respectively.

For any $x \in R^d$, several useful vector norms are given as follows,

$$\|x\|_0 = \sum_{x_i \neq 0} |x_i|^0, \quad \|x\|_p^p = \sum_{i=1}^{d} |x_i|^p, \quad \|x\|_1 = \sum_{i=1}^{d} |x_i|, \tag{1}$$

where $p \in (0, 1)$. Actually, neither l_0 nor l_p $(0 < p < 1)$ is a well defined norm because the both definitions do not satisfy the norm axioms.

$l_{2,1}$-norm of matrix was firstly introduced in [6] which can be considered as a generalization of l_1 vector norm to matrix,

$$\|A\|_{2,1} = \sum_{i=1}^{d} \|a^i\|_2. \tag{2}$$

Now we generalize the definition of $l_{2,1}$-norm to mixed $l_{2,p}$-norm as follows

$$\|A\|_{2,p} = \left(\sum_{i=1}^{d} \|a^i\|_2^p \right)^{\frac{1}{p}}, \quad p \in (0, 1]. \tag{3}$$

Note that $l_{2,p}$-norm $(0 < p < 1)$ is not a valid norm, and neither convex nor Lipschitz continuous. This properties challenge researchers to solve the related optimization problems.

Given observation data $\{a_1, a_2, \cdots, a_n\} \subseteq R^d$ and corresponding outputs $\{b_1, b_2, \cdots, b_n\} \subseteq R^c$. Traditional least square regression for discrimination solves the following optimization for unknown $X \in R^{d \times c}$

$$\min_X \sum_{i=1}^{n} \|X^T a_i - b_i\|_2^2 + \alpha R(X), \tag{4}$$

where X contains the projection matrix and bias vector for simplicity. $R(X)$ denotes regularization and $\alpha > 0$ is the regularization parameter. It is well known that the square-norm residual is sensitive to outliers, hence Nie et. al. [1] proposed to use a robust $l_{2,1}$-norm loss function

$$\min_X \sum_{i=1}^{n} \|X^T a_i - b_i\|_2 + \alpha R(X). \tag{5}$$

In this paper, we like to use the generalized version

$$\min_X \sum_{i=1}^{n} \|X^T a_i - b_i\|_2^p + \alpha R(X), \quad p \in (0, 1]. \tag{6}$$

For any $p \in (0, 1]$, the noise magnitude of distant outlier in (6) is no more than that in (5). Thus the model (6) is expected to be more robust than (5).

Joint sparse regularization $R(X)$ is usually chosen

$$R_{\triangle}(X) = \sum_{\|x^i\|_2 \neq 0}^{d} \|x^i\|_2^0 \quad \text{or} \quad R_{\triangledown}(X) = \sum_{i=1}^{d} \|x^i\|_2. \tag{7}$$

Theoretically, $R_{\triangle}(X)$ are mostly preferred for its desirable sparsity. But $R_{\triangledown}(X)$ is practically chosen more often for the computational sake. Under certain conditions, $R_{\triangledown}(X)$-regularization is equivalent to $R_{\triangle}(X)$-regularization. Here we chose the generalized norm in the sense

$$R_{\star}(X) = \sum_{i=1}^{d} \|x^i\|_2^p, \quad p \in (0, 1]. \tag{8}$$

Hence the feature selection from high-dimensional data can be concluded as a mixed optimization problem based on $l_{2,p}$-norm ($p \in (0, 1]$),

$$\min_X \sum_{i=1}^{n} \|X^T a_i - b_i\|_2^p + \gamma^p \sum_{i=1}^{d} \|x^i\|_2^p, \quad p \in (0, 1], \tag{9}$$

where $\alpha = \gamma^p$ is the regularization parameter. When $p = 1$, problem (9) is reduced to the popular $l_{2,1}$-norm based minimization proposed in [1]. But when $0 < p < 1$, it is a non-convex and non-Lipschitz continuous minimization, the algorithm in [1] can not be directly applied. As far as we know, very few scheme is presented to uniformly solve this specially mixed problem. Therefore, it is necessary to develop an unified approach to efficiently solve problem (9) for all $p \in (0, 1]$.

3 Main results

Denote $A = [a_1, a_2, \cdots, a_n] \in R^{d \times n}$ and $B = [b_1, b_2, \cdots, b_n]^T \in R^{n \times c}$, the problem (9) can be rewritten as

$$\min_X \|A^T X - B\|_{2,p}^p + \gamma^p \|X\|_{2,p}^p, \quad p \in (0, 1]. \tag{10}$$

Let $E = \frac{1}{\gamma}(A^T X - B)$, the unconstrained optimization problem (10) becomes

$$\begin{aligned} &\min_{E,X} \|E\|_{2,p}^p + \|X\|_{2,p}^p, \\ &\text{s.t.} - \gamma E + A^T X = B. \end{aligned} \tag{11}$$

It can be easily proved that $\| \begin{bmatrix} E \\ X \end{bmatrix} \|_{2,p}^p = \|E\|_{2,p}^p + \|X\|_{2,p}^p$. If we denote

$$Y := \begin{bmatrix} E \\ X \end{bmatrix} \in R^{m\times c} \quad \text{and} \quad M := [-\gamma I_n \ A^T] \in R^{n\times m}, \tag{12}$$

where $m = n + d$, problem (11) can be reformulated as

$$\begin{aligned} &\min_Y \|Y\|_{2,p}^p \\ &\text{s.t.} MY = B. \end{aligned} \tag{13}$$

The Lagrangian function of optimization problem (13) is

$$\mathcal{L}(Y, \Lambda) = \|Y\|_{2,p}^p - Tr(\Lambda^T(MY - B)). \tag{14}$$

where $\Lambda \in R^{n\times c}$ is Lagrangian multiplier matrix, and $Tr(\cdot)$ stands for trace operator. $Y^\star$ is the KKT point of problem (13) if and only if there exists a $\Lambda^\star \in R^{n\times c}$ such that

$$\begin{cases} \frac{\partial \mathcal{L}(Y,\Lambda)}{\partial Y} = pD_\star Y^\star - M^T\Lambda^\star = 0 \\ MY^\star = B \end{cases}, \tag{15}$$

where

$$D_\star = \text{diag}\{\frac{1}{\|y^1\|_2^{2-p}}, \frac{1}{\|y^2\|_2^{2-p}}, \cdots, \frac{1}{\|y^m\|_2^{2-p}}\} \tag{16}$$

is induced from $Y^\star$. After simple reformulation, (15) is equivalent to

$$Y^\star = D_\star^{-1}M^T(MD_\star^{-1}M^T)^{-1}B. \tag{17}$$

Although $D_\star$ is necessary in equation (15), only $D_\star^{-1}$ is involved to compute $Y^\star$ in formula (17). If we fix Y and D in (16) and (17) alternatively, an iterative algorithm to solve problem (13) can be designed as follows.

Algorithm 3.1 *(Solving problem (13))*

1. *Start: Given* $M \in R^{n\times m}$, $B \in R^{n\times c}$ *and set* $D_1 = I_m$.
2. *For* $k = 1, 2, \cdots$ *until convergence do :*
   $Y_{k+1} = D_k^{-1}M^T(MD_k^{-1}M^T)^{-1}B$,
   Update D_{k+1}^{-1} *with diagonal entries :* $\|y_{k+1}^i\|_2^{2-p}, i = 1, 2, \cdots, m$.

Remark 3.1 If D, Y are decided as in (16) and (17), it can be easily derived that $Tr(Y^TDY) = \|Y\|_{2,p}^p$ for $0 < p \le 1$.

Denote $\{Y_k\}$ the matrix sequence generated by Algorithm 3.1, now let us show its convergence. The first lemma is apparent so the proof is omitted.

Lemma 3.1 *If* $\phi(t) = \frac{2}{2-p}t - \frac{p}{2-p}t^{\frac{2}{p}} - 1$ *for* $p \in (0, 1]$, *then* $\phi(t) \leq 0$ *in* $(0, +\infty)$ *and* $t = 1$ *is the unique maximum point.*

Lemma 3.2 *Suppose that* y_k^i, y_{k+1}^i *is the* $i-th$ *row of* Y_k, Y_{k+1} *respectively, then for any* p *in* $(0, 1]$

$$\|y_{k+1}^i\|_2^p - \frac{p}{2}\frac{\|y_{k+1}^i\|_2^2}{\|y_k^i\|_2^{2-p}} \leq \|y_k^i\|_2^p - \frac{p}{2}\frac{\|y_k^i\|_2^2}{\|y_k^i\|_2^{2-p}}, \ i = 1, \cdots, m. \tag{18}$$

Equalities in (18) hold if and only if $\|y_{k+1}^i\|_2^p = \|y_k^i\|_2^p$ *for* $i = 1, 2, \cdots, m$.

Proof Let $t = \frac{\|y_{k+1}^i\|_2^p}{\|y_k^i\|_2^p}$ in $\phi(t)$, then $\phi(\frac{\|y_{k+1}^i\|_2^p}{\|y_k^i\|_2^p}) \leq 0$, that is

$$\frac{2}{2-p}\frac{\|y_{k+1}^i\|_2^p}{\|y_k^i\|_2^p} - \frac{p}{2-p}\frac{\|y_{k+1}^i\|_2^2}{\|y_k^i\|_2^2} - 1 \leq 0. \tag{19}$$

Note that $\|y_{k+1}^i\|_2^p = \|y_k^i\|_2^p$ for $i = 1, 2, \cdots, m$ is sufficient and necessary to let the equality in (19) happen. Multiplying the two sides of formula (19) with $(1 - \frac{p}{2})\|y_k^i\|_2^p$, we have

$$\|y_{k+1}^i\|_2^p - \frac{p}{2}\frac{\|y_{k+1}^i\|_2^2}{\|y_k^i\|_2^{2-p}} \leq (1 - \frac{p}{2})\|y_k^i\|_2^p, \tag{20}$$

which is also an equivalent formula of (18). □

Theorem 3.1 $\|Y_k\|_{2,p}^p$ *monotonically decreases with respect to iteration* k *until the matrix sequence* $\{Y_k\}$ *converges to the KKT point of problem (13).*

Proof From remark 3.1 and the construction of Algorithm 3.1, we can easily verify that Y_{k+1} is the optimal solution to

$$\begin{aligned} &\min_Y f_k(Y) := \tfrac{1}{2}Tr(Y^T D_k Y) \\ &\text{s.t. } MY = B. \end{aligned} \tag{21}$$

So we have

$$Tr(Y_{k+1}^T D_k Y_{k+1}) \leq Tr(Y_k^T D_k Y_k), \tag{22}$$

which is to say

$$\sum_{i=1}^m \frac{\|y_{k+1}^i\|_2^2}{\|y_k^i\|_2^{2-p}} \leq \sum_{i=1}^m \frac{\|y_k^i\|_2^2}{\|y_k^i\|_2^{2-p}}. \tag{23}$$

On the other hand, formula (18) in Lemma 3.2 shows

$$\sum_{i=1}^{m}\left(\|y_{k+1}^i\|_2^p - \frac{p}{2}\frac{\|y_{k+1}^i\|_2^2}{\|y_k^i\|_2^{2-p}}\right) \le \sum_{i=1}^{m}\left(\|y_k^i\|_2^p - \frac{p}{2}\frac{\|y_k^i\|_2^2}{\|y_k^i\|_2^{2-p}}\right) \tag{24}$$

Combining (23) and (24), we have

$$\sum_{i=1}^{m}\|y_{k+1}^i\|_2^p \le \sum_{i=1}^{m}\|y_k^i\|_2^p,$$

which is $\|Y_{k+1}\|_{2,p}^p \le \|Y_k\|_{2,p}^p$.

If $\|Y_{k+1}\|_{2,p}^p = \|Y_k\|_{2,p}^p$ happens for some k, since Y_{k+1} is constructed such that (22) and (23) hold, hence from (24) we easily derive

$$\sum_{i=1}^{m}\|y_{k+1}^i\|_2^p - \frac{p}{2}\frac{\|y_{k+1}^i\|_2^2}{\|y_k^i\|_2^{2-p}} = \sum_{i=1}^{m}\|y_k^i\|_2^p - \frac{p}{2}\frac{\|y_k^i\|_2^2}{\|y_k^i\|_2^{2-p}}. \tag{25}$$

Notice that (18) is valid for each $1 \le i \le m$, the equalities in Lemma 3.2 have to exist which means $\|y_{k+1}^i\|_2^p = \|y_k^i\|_2^p$ for $i = 1, 2, \cdots, m$. Hence D_{k+1}^{-1} equals D_k^{-1} in Algorithm 3.1, Y_{k+1} and D_{k+1} satisfy the stationary condition (17). Algorithm 3.1 generates a matrix sequence such that the objective function value monotonically decreases until it converges to the *KKT* matrix of problem (13). When $p = 1$, the convergence matrix is also the global minimizer of problem (13). □

Remark 3.2 To some extent, Algorithm 3.1 offers an alternative to solve l_p $(0 < p < 1)$ regularized problems when the number of columns in Y is 1.

Remark 3.3 Algorithm 3.1 is a unified approach to solve problem (13) for any $p \in (0, 1]$. This scheme provides algorithmic support to freely adapt p for better sparsity pattern in different data structures.

4 Computational details and practical algorithm

In Algorithm 3.1, each step has to compute $(MD_{k-1}^{-1}M^T)^{-1}$ which is expensive especially for high dimensional data. We notice that in the $k-$th iteration of Algorithm 3.1, Y_{k+1} solves (21) exactly. Actually, subproblem (21) is a quadratic programming with linear equality constraints, and there are varieties of efficient methods to solve it iteratively.

Now let us solve the subproblem (21) inexactly. Here we employ the gradient projection method in [14]. Suppose that Y_k has been generated as an approximate solution to the $(k-1)-$th subproblem. The next approximate matrix Y_{k+1} to the $k-$th subproblem (21) will be constructed from Y_k

$$Y_{k+1} = Y_k + \alpha_k S_k, \tag{26}$$

where α_k is the step length and S_k is the line search matrix. If only the objective function value $f_k(Y_{k+1})$ has sufficient reduction compared with $f_k(Y_k)$, the convergence will be guaranteed. Because the last approximation Y_k is feasible, that is $MY_k = B$, Y_{k+1} is feasible if and only if $MS_k = 0$. In the gradient projection method [14], S_k is chosen to be the projection of $-\nabla f_k(Y_k)$ on the null subspace of M, where

$$-\nabla f_k(Y_k) = -D_k Y_k \in R^{m\times c} \tag{27}$$

is the steepest direction matrix from Y_k. Let P denote the projection operator from $R^{m\times c}$ to $Null(M)$, then $S_k := -PD_kY_k$. Different P results in different numerical algorithm. Here we choose

$$P = I_m - M^T(M^T)^+, \tag{28}$$

where $(M^T)^+ = (MM^T)^{-1}M$. Since M^T has full column rank, $(MM^T)^{-1}$ is well defined. If we have the QR decomposition of M^T in the form

$$M^T = [Q_1 \;\; Q_2]\begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{29}$$

where $Q = [Q_1 \;\; Q_2]$ is an orthogonal matrix and $R \in R^{n\times n}$ is an invertible upper triangular matrix, then

$$P = I_m - M^T(M^T)^+ = I_m - Q_1Q_1^T. \tag{30}$$

It is easily verified that P is an orthogonal projection operator from $R^{m\times c}$ to its subspace $Null(M)$.

After the line search matrix S_k is fixed, the step length α_k can be computed by solving the following minimization

$$\min_{\alpha\geq 0} \varphi(\alpha) := Tr((Y_k + \alpha S_k)^T D_k(Y_k + \alpha S_k)). \tag{31}$$

The $\varphi(\alpha)$ can be detailedly rewritten as

$$\varphi(\alpha) = Tr(Y_k^T D_k Y_k) + 2Tr(S_k^T D_k Y_k)\alpha + Tr(S_k^T D_k S_k)\alpha^2. \tag{32}$$

Matrices D_k and P are obviously symmetric and positive semi-definite, hence

$$Tr(S_k^T D_k S_k) \geq 0\,, \quad Tr(S_k^T D_k Y_k) = -Tr((D_kY_k)^T P(D_kY_k)) \leq 0.$$

Then the optimal step length to (31) is

$$\alpha_k = -\frac{Tr(S_k^T D_k Y_k)}{Tr(S_k^T D_k S_k)} \geq 0. \tag{33}$$

The objective function value reduction of subproblem (21) can be evaluated

$$f_k(Y_{k+1}) = f_k(Y_k) - \frac{(Tr(S_k^T D_k Y_k))^2}{2Tr(S_k^T D_k S_k)}, \tag{34}$$

which is sufficient to guarantee the convergence of matrix sequence $\{Y_k\}$.

Based on the computational details (Eqs. (27)–(34)), a one-step gradient projection method for solving problem (13) can be concluded as follows.

Algorithm 4.1 *(One-step gradient projection method for problem (13))*

1. *Start: Given* $M \in R^{n\times m}$ *and* $B \in R^{n\times c}$.
2. *QR decompose* $M^T = Q_1 R$, *where* $Q_1 \in R^{m\times n}$ *and* $R \in R^{n\times n}$.
3. *Compute* $P = I_m - Q_1 Q_1^T$ *and* $Y_1 = Q_1 R^{-T} B$.
4. *For* $k = 1, 2, \cdots$ *until convergence do :*
   $D_k = diag\{\|y_k^1\|_2^{p-2}, \|y_k^2\|_2^{p-2}, \cdots, \|y_k^m\|_2^{p-2}\}$,
   $S_k = -P D_k Y_k$,
   $\alpha_k = -\frac{Tr(S_k^T D_k Y_k)}{Tr(S_k^T D_k S_k)}$,
   $Y_{k+1} = Y_k + \alpha_k S_k$.

Remark 4.1 The $k-$th iteration in Algorithm 4.1 is essentially the steepest descent method over the subspace $Null(M)$. Since M^T has full column rank, any accumulation point of $\{Y_k\}$ will be the KKT point of problem (13) [14] .

Remark 4.2 If $y_k^i = 0$ happens in some iteration, then D_k can not be well updated and Algorithm 4.1 breaks down. Here we treat D_k in two natural ways. One is setting the $i-$th diagonal element $\{D_k\}_{ii} = 0$ which can be considered as the generalized inverse of D_k^{-1}. The other one is to give a perturbation $\epsilon > 0$ such that $\{D_k\}_{ii} = (\sqrt{y_k^i (y_k^i)^T + \epsilon})^{p-2} \neq 0$.

The computational cost comparison between two algorithms can be easily analyzed as follows. With the same outer loop, Algorithm 4.1 substitutes one-step gradient projection to the matrix inverse $(M D_{k-1}^{-1} M^T)^{-1}$ in Algorithm 3.1. In each iteration, the time complexities are $\mathcal{O}(n^2 m + cmn) + \mathcal{O}(n^3)$ flops in Algorithm 3.1 and $\mathcal{O}(cm^2)$ flops in Algorithm 4.1. Here m, n, c denote the numbers of dimension, samples and selected features respectively. Except for extremely small sample data, Algorithm 4.1 will be faster than Algorithm 3.1 which is also confirmed by the CPU time comparison in the next section.

5 Experimental results

In our experiments, four public data sets in biological study are used. Brief description about all data sets is given as follows.

ALLAML is Leukemia gene microarray data, originally obtained by Golub et. al. [8]. There are 7129 genes, 72 samples containing two classes, acute lymphocytic leukemia (ALL) and acute mylogenous leukemia (AML).

Table 1 Classification error (%) of different $l_{2,p}$ matrix norms

	Top 20 features				Top 40 features			
$p =$	0.25	0.5	0.75	1	0.25	0.5	0.75	1
ALLAML	6.86	**4**	6.67	5.43	5.52	**4.1**	5.52	**4.1**
GLIOMA	**0**	**0**	**0**	2	2	**0**	**0**	2
LUNG	3.94	**1.98**	3.46	2.95	**1.46**	**1.46**	**1.46**	1.96
Pro-GE	4.9	**3.9**	6.81	5.9	8.71	**6.71**	8.71	9.71
Average	3.925	**2.47**	4.235	4.07	4.4225	**3.0675**	3.9225	4.4425

GLIOMA contains four classes, cancer glioblastomas, non-cancer glioblastomas, cancer oligodendrogliomas and non-cancer oligodendrogliomas [9]. There are total 50 samples and each class has 14, 14, 7, 15 samples respectively. Each sample has 12625 genes.

LUNG cancer data is available at [10]. There are 12533 genes and total 203 samples in two classes, malignant pleural mesothelioma and adenocarcinoma of the lung.

Prostate-GE data set has 12600 genes. It contains two classes, tumor and normal. 52 samples are tumor and 50 samples are normal. The dataset is available in [11].

Those data have high dimensional features around ten thousand but most of features are redundant even noise for classifications. Features with strong discrimination power always lie in a lower dimension subspace. It is important and necessary to select the most informative features for knowledge discovery and practical diagnosis in medicinal field. Before applied Algorithms 3.1 or 4.1 to feature selection, all data sets are performed the same preprocessing as in [7] to remove the redundant genes. Then the data sets are standardized to be zero-mean and normalized by standard deviation.

To demonstrate the effect of different $l_{2,p}$ matrix pseudo norms in feature selection, we implement typical $l_{2,p}$-norm based optimization problems for $p = 0.25, 0.5, 0.75$ and 1. Using the selected top 20, 40, 60, 80 features respectively, SVM classifiers are individually performed on four data sets with fivefold crosses. Based on Theorem 3.1, the reduction of objective function value estimates the convergence precision. To eliminate the magnitude influence of different data, we employ $\rho_k := \frac{\|Y_k\|_{2,p}^p - \|Y_{k+1}\|_{2,p}^p}{\|Y_k\|_{2,p}^p} \leq 10^{-5}$ as the stopping criterion of two algorithms. Under the same running environment, Algorithm 3.1 and 4.1 have the same classification accuracy which are reported in Tables 1, 2. The bold values indicate the best ones.

The experimental procedure indicates that four $l_{2,p}$-norm ($p = 0.25,\ 0.5,\ 0.75$ and 1) based minimizations do select different features, hence result in distinct classification performances. The parameter p in (0, 1] balances the sparsity and convexity of optimization problem (13). The closer to 0 the p is, the sparser the representation is. While p is very near to 1, the model is almost convex. The classification error comparisons show that non-convex $l_{2,p}$ ($0 < p < 1$) matrix norms provide alternatives

Table 2 Classification error (%) of different $l_{2,p}$ matrix norms

	Top 60 features				Top 80 features			
$p =$	0.25	0.5	0.75	1	0.25	0.5	0.75	1
ALLAML	6.86	**5.52**	6.86	8.29	8.57	**5.71**	8.57	8.57
GLIOMA	**2**	**2**	**2**	4	4	**2**	**2**	4
LUNG	9.33	**7.37**	8.37	10.3	**0.99**	**0.99**	1.48	1.48
Pro-GE	8.71	**6.71**	8.71	9.71	5.86	**3.95**	5.9	5.9
Average	6.725	**5.4**	6.485	8.075	4.855	**3.1625**	4.4875	4.9875

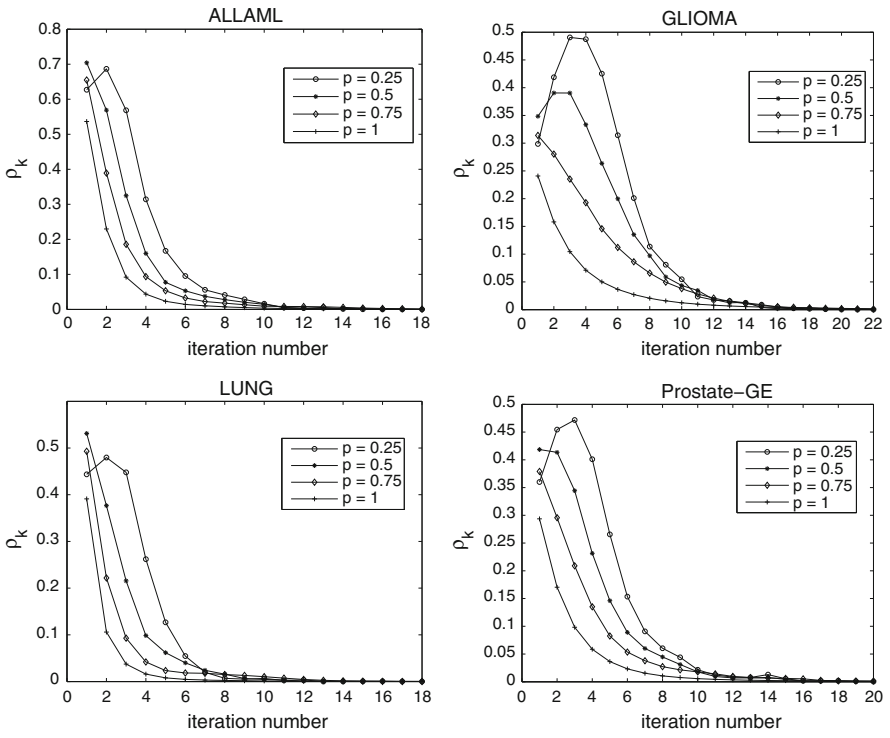

Fig. 1 The convergence performance of four $l_{2,p}$-norm based minimizations

to $l_{2,1}$-norm. Especially, $p = 0.5$ empirically outperforms $p = 1$ for choosing better sparse patterns in various situations.

To validate the consistent efficiency of the unified algorithms solving nonconvex $l_{2,p}$ $(0 < p < 1)$ pseudo norm optimization problems as well as the convex $l_{2,1}$-norm based minimization, we present the convergence behavior curve of Algorithm 3.1 with exact solution to subproblem (21). Actually, the convergence behaviors for each $l_{2,p}$-norm case are similar with different numbers of top features. We display the change of ρ_k with respect to iterations in the case of 80 features (see Fig. 1). It can be seen that

Table 3 CPU time (second) of $l_{2,p}$ matrix norms for four data sets

Data sets	ρ_k	Methods	p=0.25	p=0.5	p=0.75	p=1
ALLAML	10^{-3}	Alg 3.1	**0.75830**	0.748663	1.013861	3.235237
		Alg 4.1	0.763843	**0.652863**	**0.866831**	**2.978958**
	10^{-5}	Alg 3.1	**0.817432**	0.774233	2.646847	5.357066
		Alg 4.1	0.89743	**0.661479**	**1.894296**	**3.133529**
GLIOMA	10^{-3}	Alg 3.1	**0.502860**	**0.604722**	**0.680267**	**0.574874**
		Alg 4.1	0.596583	0.814248	0.993919	2.371908
	10^{-5}	Alg 3.1	**1.808854**	**1.752871**	**1.765150**	**3.524431**
		Alg 4.1	3.575885	1.759446	2.284880	6.803048
LUNG	10^{-3}	Alg 3.1	3.395476	3.136372	3.034811	3.524408
		Alg 4.1	**1.538439**	**1.537972**	**1.533616**	**2.392894**
	10^{-5}	Alg 3.1	3.554536	3.322220	6.279857	18.499058
		Alg 4.1	**3.081773**	**3.080179**	**5.946751**	**16.196881**
Pro-GE	10^{-3}	Alg 3.1	1.189551	1.160205	1.148347	1.501410
		Alg 4.1	**0.597630**	**0.558501**	**0.549335**	**0.630927**
	10^{-5}	Alg 3.1	1.621582	1.222292	1.794353	7.867559
		Alg 4.1	**1.120652**	**1.083130**	**1.538585**	**4.705084**

The bold values indicate the best ones

all experiments on four data sets uniformly get the expected accuracy within around 20 steps.

In practical implementation, especially for relatively large sample data, Algorithm 4.1 is preferred to Algorithm (3.1) for its economical computation. We still choose the top 80 features as a typical example to compare the efficiency of two algorithms. Under the same precisions (10^{-3} and 10^{-5}) of ρ_k, the running CPU time of two algorithms on four data sets is listed in Table 3. Algorithm 3.1 and Algorithm 4.1 are abbreviated by Alg 3.1 and Alg 4.1 respectively. All experiments are performed in the same running environment. In most of situations, Algorithm 4.1 is more time-saving than Algorithm 3.1 which especially supports Algorithm 4.1 in large scale applications.

6 Conclusions

In this paper, a type of minimizations based on $l_{2,p}$ ($p \in (0, 1]$) matrix pseudo norm is presented. A unified algorithm is designed to solve the mixed optimization problems and the convergence is also uniformly ensured. To refine the algorithm implementation, a gradient projection is applied to inexactly solve the subproblems. Experiment results on gene express data sets validate the unified performance of the proposed method. This scheme provides more choices of $p \in (0, 1]$ to fit variety of jointly sparse requirements.

Acknowledgments The first author thanks Dr. Zhang Hongchao for his helpful suggestions on this paper.

References

1. Nie, F.P., Huang, H., Cai, X., and Ding, C.: Efficient and robust feature selection via joint $l_{2,1}$-norms minimization. Twenty-Fourth Annual Conference on Neural Information Processing Systems, pp. 1–9. (2010)
2. Candès, Emmanuel J., Wakin, Michael B., Boyd, Stephen P.: Enhancing sparsity by reweighed l_1 minimization. J. Fourier Anal. Appl. **14**(5), 877–905 (2008)
3. Chartrand, R.: Exact reconstructions of sparse signals via nonconvex minimization. IEEE Signal Process. Lett. **14**(10), 707–710 (2007)
4. Chartrand, R., and Yin, W.: Iteratively reweighed algorithms for compressive sensing. 33rd International Conference on Acoustics, Speech, and Signal Processing, pp. 3869–3872 (2008)
5. Chen, X.J., Xu, F.M., Ye, Y.Y.: Lower bound theory of nonzero entries in solutions of $l_2 - l_p$ minimization. SIAM J. Sci. Comput. **32**(5), 2832–2852 (2010)
6. Ding, C., Zhou, D., He, X.F., and Zha, H.Y.: $R1-$PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization. Proceedings of the 23th International Conference on Machine Learning, pp. 281–288 (2006)
7. Dudoit, S., Fridly, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc. **97**(457), 77–87 (2002)
8. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**(5439), 531–537 (1999)
9. Nutt, C.L., Mani, D.R., Betensky, R.A., Tamayo, P., Caincross, J.G., Ladd, C., Pohl, U., Hartmann, C., Mclaughlin, M.E.: Gene expression-based classification of malignant gliomas correlates better with servival than histological classification. Cancer Res. **63**, 1602–1607 (2003)
10. Gordon, G.J., Jensen, R.V., Hsiao, L.L., Gullans, S.R., Blumenstock, J.E., Ramaswamy, S., Richards, W.G., Sugarbaker, D.J., Bueno, R.: Translation of microarray data into clinically relevant cancer diagnoistic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res. **62**(17), 4963–4967 (2002)
11. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., Sellers, W.R.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell **1**(2), 203–209 (2002)
12. Xu, Z.B., Zhang, H., Wang, Y., Chang, X.Y., Yong, L.: $L_{\frac{1}{2}}$ regularizer. Sci. China **52**(6), 1159–1169 (2010)
13. Rakotomamonjy, A., Flamary, R., Gasso, G., Canu, S.: $l_p - l_q$ Penalty for sparse linear and sparse multiple kernel multitask learning. IEEE Transac. Neural Netw. **22**(8), 1307–1320 (2011)
14. Rosen, J.B.: The gradient projection method for nonlinear programming. Part 1 Linear constraints. J. SIAM **8**, 181–217 (1960)