Communications in Information and Systems
Volume 14, Number 2, 87–109, 2014

Ratio and difference of l_1 and l_2 norms and sparse representation with coherent dictionaries

Penghang Yin, Ernie Esser, and Jack Xin

The ratio of l_1 and l_2 norms has been used empirically to enforce sparsity of scale invariant solutions in non-convex blind source separation problems such as nonnegative matrix factorization and blind deblurring. In this paper, we study the mathematical theory of the sparsity promoting properties of the ratio metric in the context of basis pursuit via over-complete dictionaries. Due to the coherence in the dictionary elements, convex relaxations such as l_1 minimization or non-negative least squares may not find the sparsest solutions. We found sufficient conditions on the nonnegative solutions of the basis pursuit problem so that the sparsest solutions can be recovered exactly by minimizing the nonconvex ratio penalty. Similar results hold for the difference of l_1 and l_2 norms. In the unconstrained form of the basis pursuit problem, these penalties are robust and help select sparse if not the sparsest solutions. We give analytical and numerical examples and introduce sequentially convex algorithms to illustrate how the ratio and difference penalties are computed to produce both stable and sparse solutions.

AMS 2000 subject classifications: 94A12, 94A15, 90C26, 90C25.
Keywords and phrases: l_1 and l_2 norms, ratio and difference, coherent dictionary, sparse representation.

1. Introduction

The ratio of l_1 and l_2 norms is a widely used empirical nonconvex scale-invariant penalty for encouraging sparsity for nonconvex problems such as nonnegative matrix factorization (NMF) and blind deconvolution applications [11, 12, 13]. A related metric that is homogeneous of degree one is the difference of l_1 and l_2 norms. Both appear to encourage sparse solutions to non-negative least squares (NNLS) type problems of the form

$$\min_{x\geq 0} \frac{\lambda}{2}\|A\,x-b\|^2 + R(x)\ , \tag{1.1}$$

where $R(x) = \frac{\|x\|_1}{\|x\|_2}$ or $R(x) = \|x\|_1 - \|x\|_2$. If A satisfies certain incoherence properties, then sufficiently sparse nonnegative solutions to $A\,x = b$ are unique [1], which is why solving the convex NNLS problem often works well without the additional sparsity penalty $R(x)$. A coherence measure of matrix A is $\rho(A)$ defined as the maximum of the cosine of pairwise angles between any two columns of A. Let $t_A = \frac{\rho}{1+\rho}$, as in [1]. A sufficient incoherence condition [1] for uniqueness of the sparsest solution $x_0 \geq 0$ is that $\|x_0\|_0 < \frac{1}{2t_A}$. However, such conditions are often not satisfied in practice, in which case including $R(x)$ can yield much sparser solutions. Likewise, minimizing the convex metric, the l_1 norm, can effectively recover sparse solutions to the underdetermined system $A\,x = b$ when columns of the matrix A satisfy certain incoherence conditions [2, 3, 4].

We are interested in understanding the sparsity promoting properties of the ratio and difference metrics theoretically and computationally in a highly coherent (over-complete) dictionary. Over-complete dictionaries occur in human visual and auditory systems [14, 10, 16], and in discretization of continuum imaging problems such as radar and medical imaging when the grid spacing is below the Rayleigh threshold [7]. Band exclusion and local optimization techniques are introduced to image objects sufficiently separated with respect to the coherence bands in [7].

Computationally, the two non-convex penalties can be treated as follows. Since $\|x\|_1 - \|x\|_2$ is a difference of convex functions, stationary points of the resulting nonconvex model are computed by difference of convex (DC) programming [15]. The model with the ratio penalty $\frac{\|x\|_1}{\|x\|_2}$ can be minimized using a related gradient projection strategy. In the general dictionary case, the exact l_1 recovery of sparse solutions is studied in [5] where a main result is that if $A\,x_0 = f$, $\|x_0\|_0 < \frac{1+M^{-1}}{2}$, M being an upper bound of off diagonal entries of the Gram matrix $A^T A$, then x_0 is the unique solution given by l_1 norm minimization. The columns of A are more coherent if M is larger. In this case, l_1 minimization is less effective.

The organization of the paper is as follows. In section 2, we begin with examples of the basis pursuit problem of the form $\min_x R(x)$ subject to $A\,x = b$ to compare l_1 or l_p ($p \in (0, 1)$) minimization with that of the ratio or difference of l_1 and l_2 norms, and with the ground truth to understand the properties and limitations of each metric. These analytical examples help to introduce the coherence issues in finding sparse solutions. We leave as a future work to investigate similar phenomena in physical data sets. In section 3, we show that mimimizers of the ratio or difference of l_1 and l_2 norms must be locally the sparsest feasible solution. We then formulate a uniformity condition on a particular subset $\mathcal{F}_L$ of the feasible solutions and prove the exact recovery of the sparsest solution x_0 of $A\,x = b$ by minimizing the ratio of l_1 and l_2 norms. The uniformity condition essentially says that the ratio of the minimum and maximum of the nonzero entries of any solution x from $\mathcal{F}_L$ is bounded from below by a constant that depends on $\|x_0\|_0$ and $\|x\|_0$. Interestingly, a similar condition appears in [7] for the band-excluded orthogonal matching pursuit method to recover the support of the solution up to the coherence band. The ratio of the maximum and minimum over the support of a vector is referred to as dynamic range in optical imaging [7]. For the difference of l_1 and l_2 norms, the exact recovery condition is that the minimum of the nonzero entries of any solution $x \neq x_0$ from $\mathcal{F}_L$ be above $\frac{2(\sqrt{\|x_0\|_0}-1)}{\|x\|_0-1}\,\|x_0\|_2$. Our theoretical results shed light on the sparsity promoting capability of the ratio and difference penalties. In section 4, we show numerical examples optimizing (1.1) with A being a coherent dictionary such that the ratio and difference of l_1 and l_2 norms regularization outperform NNLS. More comparisons with l_1 minimization for imaging data can be found in [6]. Concluding remarks are in section 5.

2. Examples of basis pursuit in coherent dictionaries

The examples below will show a couple of situations where l_1 or l_p ($p \in (0,1)$) minimization ceases to be effective. Though these examples are mathematical in nature, they help to illustrate the coherence induced issues in finding sparse solutions. Imaging examples can be found in [7, 6] where the sparse solutions are more complicated and not in closed analytical form.

Example 1: Let $p \in (0,1]$ and two distinct dense vectors b^1, $b^2 \in \mathbb{R}^n$ ($n \geq 2$). so that $b = b^1 + b^2$ is also dense; Let $\frac{\|b^i\|_1}{\|b^i\|_2}$ be close to their upper bound $O(\sqrt{n})$, $i = 1, 2$. $a = \|(b^1, b^2)\|_p$, $A = [b^1, b^2, a\,I_n, a\,I_n]$, where I_n is $n \times n$ identity matrix. Consider the linear system $A\,x = b$, $x \in \mathbb{R}^{2+2n}$, which has a 2-sparse solution:

$$x_0 = [1, 1, 0, \cdots, 0]'.$$

The other sparse solutions are: $x_1 = [0, 1, \frac{(b^1)'}{a}, 0]'$, $x_2 = [1, 0, 0, \frac{(b^2)'}{a}]'$, $x_3 = [0, 0, \frac{(b^1)'}{a}, \frac{(b^2)'}{a}]'$, the first two are at least 3-sparse, the last one is at least 4-sparse. The l_p norm of x_s is:

$$\|x_0\|_p = 2^{1/p},$$

while:

$$\|x_1\|_p = (1 + \frac{\|b^1\|_p^p}{a^p})^{1/p} \in (1, 2^{1/p}),$$
$$\|x_2\|_p = (1 + \frac{\|b^2\|_p^p}{a^p})^{1/p} \in (1, 2^{1/p}),$$
$$\|x_3\|_p = \frac{\|[(b^1)', (b^2)']'\|_p}{a} = 1.$$

Thus, x_0 cannot be recovered by minimizing l_p norm subject to $Ax = b$. There are at least three less sparse solutions with smaller l_p norm than x_0.

Now let $p = 1$, the l_2 norms of x_0 and x_3 are:

$$\|x_0\|_2 = \sqrt{2},\ \|x_3\|_2 = \frac{\|[(b^1)', (b^2)']'\|_2}{\|[(b^1)', (b^2)']'\|_1}.$$

So the ratio of l_1 and l_2 norms are:

$$\frac{\|x_0\|_1}{\|x_0\|_2} = \sqrt{2},\ \frac{\|x_1\|_1}{\|x_1\|_2} = \frac{\|[a', (b^1)']'\|_1}{\|[a', (b^1)']'\|_2},$$
$$\frac{\|x_2\|_1}{\|x_2\|_2} = \frac{\|[a', (b^2)']'\|_1}{\|[a', (b^2)']'\|_2},\ \frac{\|x_3\|_1}{\|x_3\|_2} = \frac{\|[(b^1)', (b^2)']'\|_1}{\|[(b^1)', (b^2)']'\|_2} \sim \sqrt{n}.$$

We want to have $\frac{\|x_1\|_1}{\|x_1\|_2} > \sqrt{2}$ or:

$$\frac{\|[a', (b^1)']'\|_1}{\|[a', (b^1)']'\|_2} = \frac{2\|b^1\|_1 + \|b^2\|_1}{\sqrt{(\|b^1\|_1 + \|b^2\|_1)^2 + \|b^1\|_2^2}} > \sqrt{2},$$

or:

$$2\|b^1\|_1^2 > \|b^2\|_1^2 + 2\|b^1\|_2^2.$$

Likewise $\frac{\|x_2\|_1}{\|x_2\|_2} > \sqrt{2}$ requires:

$$2\|b^2\|_1^2 > \|b^1\|_1^2 + 2\|b^2\|_2^2.$$

The above inequalities reduce to:

$$\|b^i\|_1 > \sqrt{2}\|b^i\|_2,\ \ i = 1, 2,$$

if we assume that the first two columns of A satisfy $\|b^1\|_1 = \|b^2\|_1$, $b^1 \neq b^2$. It follows that x_0 has the smallest ratio of l_1 and l_2 norms. So $\frac{l_1}{l_2}$ minimization can recover x_0 in the counterexample where l_1 minimization could not.

Let us look at difference of l_1 and l_2 norms at $p = 1$.

$$\|x_0\|_1 - \|x_0\|_2 = 2 - \sqrt{2}.$$

$$\|x_3\|_1 - \|x_3\|_2 = 1 - \frac{\|[(b^1)', (b^2)']'\|_2}{\|[(b^1)', (b^2)']'\|_1} = 1 - O(n^{-1/2}) > 2 - \sqrt{2},$$

if n is large enough. However,

$$\|x_1\|_1 - \|x_1\|_2 = 1 + \frac{\|b^1\|_1}{a} - \sqrt{1 + \frac{\|b^1\|_2^2}{a^2}}.$$

$$\|x_2\|_1 - \|x_2\|_2 = 1 + \frac{\|b^2\|_1}{a} - \sqrt{1 + \frac{\|b^2\|_2^2}{a^2}}.$$

If both were above $2 - \sqrt{2}$ so that x_s has the least difference of l_1 and l_2 norms, we would have by adding the two expressions:

$$4 - 2\sqrt{2} \leq 3 - \sum_{i=0,1} \sqrt{1 + \frac{\|b^i\|_2^2}{\|[(b^1)', (b^2)']'\|_1^2}} \leq 3 - \sum_{i=0,1} 1,$$

or:

$$4 - 2\sqrt{2} \approx 1.1716 \leq 1,$$

which is impossible. Hence minimizing the difference of l_1 and l_2 norms gives either x_1 or x_2, the 2nd sparsest solution, but not the sparsest solution x_0 in this example. It is better than minimizing l_1 which gives the 3rd sparsest solution x_3.

Since the $\frac{l_1}{l_2}$ penalty tends to get larger for more dense vectors, it is plausible that x_0 is recovered by minimizing $\frac{l_1}{l_2}$ if n is large enough. However, this cannot happen without proper conditions on b^1, b^2. We show a counterexample below.

First, we note that

$$\mathbf{Ker}(A) = \mathbf{span}\{[1, 0, -\frac{(b^1)'}{a}, 0]', [0, 1, -\frac{(b^2)'}{a}, 0]', [0, 0, -c', c']'\},\ \forall\, c \in \mathbb{R}^n.$$

Let

$$x_4 = x_0 + [1, 0, -\frac{(b^1)'}{a}, 0]' - [0, 1, \frac{(b^2)'}{a}, 0]' = [2, 0, \frac{(b^2 - b^1)'}{a}, 0]'.$$

Then

$$\frac{\|x_4\|_1}{\|x_4\|_2} \leq \frac{2 + \frac{\|b^2 - b^1\|_1}{a}}{2} < \frac{\|x_0\|_1}{\|x_0\|_2} = \sqrt{2},$$

if

$$\frac{\|b^2 - b^1\|_1}{a} < 2\sqrt{2} - 2 \approx 0.828.$$

Since $\frac{\|b^2-b^1\|_1}{a} \leq \frac{\|b^2\|_1+\|b^1\|_1}{a} = 1$, this is not a stringent condition. Thus x_4 is a less sparse solution than x_0 with smaller ratio of $\frac{l_1}{l_2}$ norm. Minimization of $\frac{l_1}{l_2}$ does not yield x_0. On the other hand, x_4 contains a large peak (height 2), and many smaller peaks ($\frac{b^1-b^2}{a}$), resembling a perturbation of the 1-sparse solution $[2, 0, \cdots, 0]'$ in the case of $b^1 = b^2$.

In particular, if b^2 is a small perturbation of b^1, then $\frac{\|b^2-b^1\|_1}{a} \approx 0$. So x_4 is close to the 1-sparse vector $[2, 0, \cdots, 0]'$ with the ratio of $\frac{l_1}{l_2}$ norm slightly above 1, the least value of the ratio among all nonzero vectors. We observe here that $\min_{x:Ax=b} \frac{\|x\|_1}{\|x\|_2}$ is continuous with respect to the perturbations of A. The minimizer goes from exact 1-sparse structure when $b^1 = b^2$ to an approximate 1-sparse structure when $b^1 \approx b^2$. In contrast, the l_0 minimizer x_0 experiences a jump from $[2, 0, 0, 0]'$ to $[1, 1, 0, 0]'$. The discrete character of l_0 makes it non-trival to recover the least l_0 solution from minimizing $\frac{l_1}{l_2}$. If we view b^1 and b^2 as dictionary elements in a group, then minimizing $\frac{l_1}{l_2}$ selects only one of them (intra sparsity). Similarly, if we view corresponding columns (1st and $(n+1)$-th, 2nd and $(n+2)$-th, etc) of $[\alpha I_n \;\; \alpha I_n]$ as vectors in a group (of 2 elements), then x_4 selects one member out of each group. The examples here show that minimizing $\frac{l_1}{l_2}$ has the tendency of removing redundencies or preferring intra-sparsity in a coherent and over-determined dictionary. The l_1 minimization does not do as well in terms of intra-sparsity, using all group elements except for knocking out the b^1, b^2 group.

Let us look closer at the solutions in the non-negative orthant. Such vectors are:

$$x = [1 + t_1, 1 + t_2, \; -(t_1\frac{b^1}{a} - c)', \; -(t_2\frac{b^2}{a} + c)']',$$

satisfying:

$$1 + t_1 \geq 0, \; 1 + t_2 \geq 0, \; t_2\frac{b^2}{a} \leq c \leq -t_1\frac{b^1}{a},$$

which is valid if:

$$b^1 < 0, \; t_1 \in (0, \frac{2}{3}), \; b^2 = (1-\epsilon)b^1, \; 0 < \epsilon \ll 1, \; t_2 \approx -t_1. \tag{2.2}$$

The kernel is an $(n+2)$-dimensional plane which contains a lower dimensional affine subspace parallel to the unit l_1 ball if the vectors on the plane:

$$v = [t_1, t_2, -(t_1\frac{b^1}{a} - c)', -(t_2\frac{b^2}{a} + c)']',$$

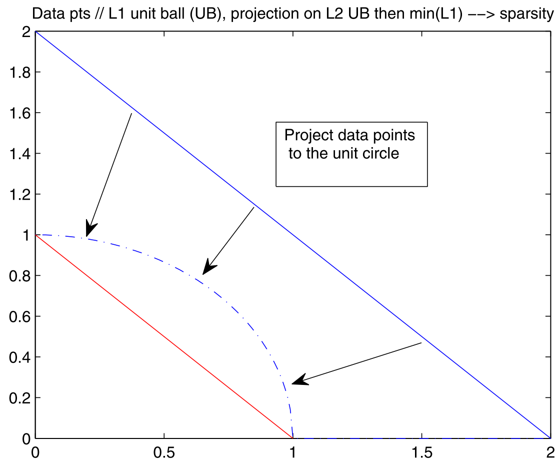

Figure 1: Illustration of the advantage of minimizing $\frac{l_1}{l_2}$ over l_1 when data points (on $x_1 + x_2 = 2$ in the first quadrant) lie parallel to the l_1 unit ball. Minimization of $\frac{l_1}{l_2}$ is same as projection onto the unit l_2 ball then intersecting with the unit l_1 ball to select sparse solutions. In contrast, l_1 minimization cannot distinguish sparse data points.

are orthogonal to the one vector $[1, 1, \cdots, 1, 1]' \in \mathbb{R}^{2n+2}$, in other words,

$$\sum_{i=1,2} \left(1 - \sum_j \frac{b^i_j}{a}\right) t_i = 0, \tag{2.3}$$

which holds with essentially an $(n+1)$-dimensional free parameter (t_1, c), under constraints in (2.2). If the minimal l_1 ball intersects the kernel at a point p, the line at p in the direction v lies on the l_1 ball. Then l_1 minimization is not effective, there are infinitely many non-sparse minimizers. An illustration in two dimensions is in Fig. 1, where all points on $x_1 + x_2 = 2$ in the first quadrant are minimizers of l_1 norm. Using the scale invariance of $\frac{l_1}{l_2}$, minimizing $\frac{l_1}{l_2}$ can be viewed as first projecting data points (feasible vectors) onto the l_2 unit ball, then intersecting with the minimal l_1 ball, which leads to sparse solutions.

Example 2: Let $p \in (0, 1]$ and a dense vector $b \in \mathbb{R}^n$ $(n \geq 2)$, $a = \|(b, b)\|_p = 2^{1/p}\|b\|_p$, $A = [b, b, a\, I_n, a\, I_n]$, where I_n is $n \times n$ identity matrix. The linear system $A\, x = 2b$, $x \in \mathbb{R}^{2+2n}$ has a 1-sparse solution:

$$x_0 = [2, 0, \cdots, 0]'.$$

There is also a 2-sparse solution

$$x_1 = [1, 1, 0, \cdots, 0]'.$$

Some other solutions are: $x_2 = [1, 0, \frac{b'}{a}, 0]'$, $x_3 = [0, 0, \frac{b'}{a}, \frac{b'}{a}]'$, Then minimizing $\frac{l_1}{l_2}$ and $l_1 - l_2$ both give the sparsest solution x_0, since both $\frac{\|x_0\|_1}{\|x_0\|_2}$ and $\|x_0\|_1 - \|x_0\|_2$ attain their possible lower bounds 1 and 0. However, for l_p-norm minimization ($p \in (0, 1]$),

$$\begin{aligned}
&\|x_0\|_p = 2, \\
&\|x_1\|_p = 2^{1/p} \geq 2, \\
&\|x_2\|_p = (1 + \frac{\|b\|_p^p}{a^p})^{1/p} = (\frac{3}{2})^{1/p} \geq \frac{3}{2}, \\
&\|x_3\|_p = \frac{\|[b', b']'\|_p}{a} = 1.
\end{aligned}$$

For $p = 1$, x_0 has the largest l_1 norm among these solutions. For $p \in (0, 1]$, $\|x_0\|_p > \|x_3\|_p$. So l_p-norm minimization fails to find the sparsest solution.

Example 3: Let $A = [b^1, b^2, I_n, a\, I_n]$ be the same from Example 1 and $b = b^1 + e^1$, where $e^1 = [1, 0, \cdots, 0]'$, $a = 2^{1/p}\|b^1\|_p$ ($p \in (0, 1]$), $b^2 \neq b^1$, both dense. The aim is to represent data b with columns of A to have both intra-sparsity and inter-sparsity across the groups.

The 2-sparse solutions with perfect intra and inter sparsity (at most 1 in each group and least number of groups) are:

$$x_0^1 = [1, 0, (e^1)', 0]', \quad x_0^2 = [1, 0, 0, \frac{(e^1)'}{a}]',$$

some much less sparse solutions are (good intra-sparsity, almost no inter-sparsity):

$$\begin{aligned}
x_1 &= [0, 0, (e^1)', \frac{(b^1)'}{a}]', \\
x_2 &= [0, 0, 0, \frac{(e^1 + b^1)'}{a}]'.
\end{aligned}$$

We have:

$$\|x_0^1\|_p^p = 2 > 1 + \frac{1}{2} = \|x_1\|_p^p$$

$$\|x_0^2\|_p^p = 1 + \frac{1}{a^p} > \frac{1}{2} + \frac{1}{a^p} \geq \|x_2\|_p^p.$$

So l_p minimization will miss the 2-sparse solutions.

Let $p = 1$, in view of:

$$\frac{\|x_0^1\|_1}{\|x_0^1\|_2} = \sqrt{2} \approx 1.414,$$
$$\frac{\|x_0^2\|_1}{\|x_0^2\|_2} = \frac{1 + a^{-1}}{\sqrt{1 + a^{-2}}} \leq \sqrt{2},$$
$$\frac{\|x_1\|_1}{\|x_1\|_2} = \frac{1.5}{\sqrt{1 + \frac{\|b^1\|_2^2}{4\|b^1\|_1^2}}} \approx 1.5^-,$$

if $\|b^1\|_1 \gg \|b^1\|_2$ by the assumption, $\frac{l_1}{l_2}$ of x_0^1 or x_0^2 can be smaller. If a is large, $\frac{l_1}{l_2}$ minimization prefers x_0^2 because it is a small perturbation of a 1-sparse vector $[1, 0, 0, 0]'$. However, minimizing $\frac{l_1}{l_2}$ does not always lead to x_0^2 if a is small enough. We show a counterexample as below: Let $x_3 = [0, 0, (b^1)', \frac{(e^1)'}{a}]'$, then

$$\frac{\|x_3\|_1}{\|x_3\|_2} \leq \frac{\frac{a}{2} + a^{-1}}{a^{-1}} < \frac{\|x_0^2\|_1}{\|x_0^2\|_2} = \frac{1 + a^{-1}}{\sqrt{1 + a^{-2}}},$$

if

$$a < 0.908.$$

Notice that x_3 has one large peak and many (relatively speaking) smaller peaks, resembling a small perturbation of a 1-sparse vector if a is small enough.

The solutions are of the form:

$$x = [1, 0, 0, \frac{(e^1)'}{a}]' + t_1[1, 0, -(b^1)', 0]' + t_2[0, 1, -(b^2)', 0]' + [0, 0, a\, c', c']',$$

nonnegativity constraints are:

$$1 + t_1 \geq 0,\ t_2 \geq 0,\ -t_1 b^1 - t_2 b^2 + a\, c \geq 0, \frac{e^1}{a} + c \geq 0. \tag{2.4}$$

In particular, consider

$$t_2 \geq 0,\ c \geq 0.$$

At any point p on the plane $A\,x = b$, we seek a direction

$$v = t_1[1, 0, -(b^1)', 0]' + t_2[0, 1, -(b^2)', 0]' + [0, 0, a\,c', c']',$$

so that $v \cdot [1, \cdots, 1] = 0$, or:

$$t_1 + t_2 + \sum_j \; -t_1\, b^1_j - t_2\, b^1_j + (a+1)c_j = 0,$$

or:

$$(1 - \sum_j b^1_j)\, t_1 + (1 - \sum_j b^2_j)\, x_2 + (a+1) \sum_j c_j = 0,$$

which admits nontrivial solutions satisfying (2.4) if

$$\begin{aligned}
&c = 0,\ \sum_j b^1_j = 0,\ \sum_j b^2_j = 0,\\
&t_1 = -t_2,\ t_2 \in (0,1),\\
&-t_1 b^1 - t_2 b^2 = t_2(b^1 - b^2) > 0.
\end{aligned}$$

So the intersection of the l_1 minimal ball with the kernel is at least a line segment, rendering l_1 minimizers non-unique and most of the l_1 minimizers non-sparse.

In summary, the examples here indicate that minimizing the ratio of l_1 and l_2 norms is more likely to get a sparser solution than minimizing l_p ($p \in (0,1)$) when the column vectors of matrix A are structured or coherent. The geometric reason is that the l_1 unit ball with corners and edges tend to hit the unit sphere on axes or coordinate planes resulting in sparse solutions. Intersecting the l_1 unit ball with another high dimensional plane may have multiple non-sparse minimizers (as shown in Fig. 1). Minimizing the difference of l_1 and l_2 norms is better than minimizing l_p norms, and appears no better than minimizing the ratio of l_1 and l_2 norms. Computationally though, the difference has better analytical structure for algorithm design as we shall explore later.

3. Exact recovery theory

In this section, we show that it is possible to recover the sparsest solution exactly by minimizing the ratio and difference of l_1 and l_2 norms, thereby establishing the origin of their sparsity promoting property.

3.1. Exact recovery of $\frac{l_1}{l_2}$

Suppose $A \in \mathbb{R}^{m\times n}$ and $x_0 \geq 0 \in \mathbb{R}^n$, where $m < n$. Let $b = A\,x_0$, we exclude the case $b = 0$ throughout this paper and study the following problems:

$$\begin{aligned}
P_0: &\quad \min_{x\geq 0} \|x\|_0 \quad \text{subject to} \quad A\,x = b\\
P_r: &\quad \min_{x\geq 0} \frac{\|x\|_1}{\|x\|_2} \quad \text{subject to} \quad A\,x = b\\
P_d: &\quad \min_{x\geq 0} \|x\|_1 - \|x\|_2 \quad \text{subject to} \quad A\,x = b
\end{aligned}$$

Denote by $\mathcal{F} = \{x \in \mathbb{R}^n : A\,x = b,\ x \geq 0\}$ the set of feasible solutions, and let $S(x)$ denote the support of x.

Definition 3.1. *$x \in \mathcal{F}$ is called* **locally sparse** *if $\nexists\, y \in \mathcal{F}\setminus\{x\}$ such that $S(y) \subseteq S(x)$. Denote by $\mathcal{F}_L = \{x \in \mathcal{F} : x \text{ is locally sparse}\}$ the set of locally sparse feasible solutions.*

The following lemma says that any locally sparse solution is in essence locally the sparsest solution.

Lemma 3.1. *$\forall\, x \in \mathcal{F}_L$, $\exists\, \delta_x > 0$ such that $\forall\, y \in \mathcal{F}$, if $0 < \|y - x\|_2 < \delta_x$, we have $S(x) \subset S(y)$.*

Proof. Let $y = x + v$ and choose $\delta_x = \min_{i\in S(x)}\{x_i\}$, then

$$\|v\|_\infty \leq \|v\|_2 < \min_{i\in S(x)}\{x_i\}$$

So

$$y_i \geq x_i - \|v\|_\infty > x_i - \min_{i\in S(x)}\{x_i\} \geq 0, \forall\, i \in S(x)$$

which implies

$$S(x) \subseteq S(y).$$

And $S(x) \neq S(y)$ since $x \in \mathcal{F}_L$. Then the claim follows. $\square$

The following theorem states that the solutions of P_r, P_d and P_0 must be locally sparse, thereby being at least locally the sparsest feasible solution.

Theorem 3.1. *If x^* solves P_r, P_d or P_0, then $x^* \in \mathcal{F}_L$.*

Proof. Suppose x^* solves P_r or P_d and it is not locally sparse, then $\exists\, y^* \in \mathcal{F} \setminus \{x^*\}$ such that $S(y^*) \subseteq S(x^*)$. Thus there exists a small enough $\epsilon > 0$, such that $x^* - \epsilon y^* \geq 0$. Let

$$z^* = \frac{x^* - \epsilon y^*}{1 - \epsilon} \geq 0$$

or equivalently,

$$x^* = \epsilon y^* + (1 - \epsilon) z^*$$

then $A\, z^* = b$ and thus $z^* \in \mathcal{F}$.

By the nonnegativity of y^* and z^*,

$$\|x^*\|_1 = \epsilon \|y^*\|_1 + (1 - \epsilon)\|z^*\|_1$$

Moreover, since $y^* \neq x^*$ both satisfying $A\, x = b$, they are linearly independent. So y^* and z^* are also linearly independent, and

$$\|x^*\|_2 < \epsilon \|y^*\|_2 + (1 - \epsilon)\|z^*\|_2$$

Thus

$$\frac{\|x^*\|_1}{\|x^*\|_2} > \frac{\epsilon \|y^*\|_1 + (1 - \epsilon)\|z^*\|_1}{\epsilon \|y^*\|_2 + (1 - \epsilon)\|z^*\|_2} \geq \min\{\frac{\|y^*\|_1}{\|y^*\|_2}, \frac{\|z^*\|_1}{\|z^*\|_2}\}$$

and

$$\begin{aligned} \|x^*\|_1 - \|x^*\|_2 &> \epsilon(\|y^*\|_1 - \|y^*\|_2) + (1 - \epsilon)(\|z^*\|_1 - \|z^*\|_2) \\ &\geq \min\{\|y^*\|_1 - \|y^*\|_2, \|z^*\|_1 - \|z^*\|_2\} \end{aligned}$$

Contradiction.

Now suppose x^* solves P_0 and it is not in $\mathcal{F}_L$, then $\exists\, y^* \in \mathcal{F} \setminus \{x^*\}$ such that $S(y^*) \subseteq S(x^*)$. Since no nonnegative solution of $A\, x = b$ is sparser than x^*, we have $S(y^*) = S(x^*) = S$. So $\min_{i \in S}\{\frac{x_i^*}{y_i^*}\} < 1$ or $\min_{i \in S}\{\frac{y_i^*}{x_i^*}\} < 1$ must be true. Without loss of generality, let $\min_{i \in S}\{\frac{x_i^*}{y_i^*}\} = \frac{x_k^*}{y_k^*} = r < 1$ for some index $k \in S$. Then $z^* = \frac{1}{1-r} x^* - \frac{r}{1-r} y^* \geq 0$ since $z_i^* = \frac{x_i^* - r y_i^*}{1-r} \geq 0$, $\forall\, i \in S$. Moreover, $A\, z^* = \frac{1}{1-r} A\, x^* - \frac{r}{1-r} A\, y^* = \frac{1}{1-r} b - \frac{r}{1-r} b = b$ which implies $z^* \in \mathcal{F}$. But $S(z^*) \subseteq S$ and $z_k^* = 0$, thus $S(z^*) \subset S$ which contradicts with x^* being the solution of P_0. ☐

By Theorem 3.1, all the minimizers of P_r, P_d and P_0 are contained in $\mathcal{F}_L$. From now on, we no longer care about those feasible solutions outside $\mathcal{F}_L$.

For any $x \geq 0 \in \mathbb{R}^n$, suppose $(S(x), Z(x))$ is a partition of the index set of x, i.e., $\{1, 2, \cdots, n\}$, where $S(x) = \{i : x_i > 0\}, Z(x) = \{i : x_i = 0\}$.

Definition 3.2. *The uniformity of x, U(x), is the ratio between the smallest nonzero entry and the largest one, i.e.*

$$0 < U(x) := \frac{\min_{i \in S(x)} x_i}{\max_{i \in S(x)} x_i} \leq 1.$$

Theorem 3.2. *If x_0 uniquely solves P_0 and $\|x_0\|_0 = s$, if $U(x) > \frac{\sqrt{\|x\|_0} - \sqrt{\|x\|_0 - s}}{\sqrt{\|x\|_0} + \sqrt{\|x\|_0 - s}}$, $\forall\, x \in \mathcal{F}_L \setminus \{x_0\}$, then x_0 also uniquely solves P_r. In particular, if any feasible solution x is a binary vector with all entries either 0 or 1, then the above inequality holds since $U(x) = 1 > \frac{\sqrt{\|x\|_0} - \sqrt{\|x\|_0 - s}}{\sqrt{\|x\|_0} + \sqrt{\|x\|_0 - s}}$. Clearly P_0 and P_r are equivalent as we note that $\frac{\|x\|_1}{\|x\|_2} = \sqrt{\|x\|_0}$.*

Since $\frac{\|x\|_1}{\|x\|_2}$ is scale-invariant, $\forall x \geq 0 \in \mathbb{R}^n$, without loss of generality, we assume $\max_{i \in S(x)} x_i = 1$ and $0 < \min_{i \in S(x)} x_i = U(x) \leq 1$. By the Cauchy-Schwarz inequality, $\frac{\|x\|_1}{\|x\|_2} \leq \sqrt{\|x\|_0}$. Starting with the following lemma, we first estimate the lower bound of $\frac{\|x\|_1}{\|x\|_2}$.

Lemma 3.2. *Let $x = [x_1, \cdots, x_{j-1}, x_j, x_{j+1}, \cdots, x_n]'$, where $U(x) < x_j < 1$.*
Let $x_- = [x_1, \cdots, x_{j-1}, U(x), x_{j+1}, \cdots, x_n]'$ and $x_+ = [x_1, \cdots, x_{j-1}, 1, x_{j+1}, \cdots, x_n]'$, then we have

$$\frac{\|x\|_1}{\|x\|_2} > \min\{\frac{\|x_-\|_1}{\|x_-\|_2}, \frac{\|x_+\|_1}{\|x_+\|_2}\}$$

Proof. Since $U(x) < x_j < 1$, $\exists\, \lambda > 0$, such that $x_j = \lambda U(x) + (1-\lambda)1$ and $x = \lambda x_- + (1-\lambda)x_+$. Given that x_- and x_+ are nonnegative but not linearly dependent, we have

$$\|x\|_1 = \|\lambda x_- + (1-\lambda)x_+\|_1 = \lambda\|x_-\|_1 + (1-\lambda)\|x_+\|_1$$

and

$$\|x\|_2 = \|\lambda x_- + (1-\lambda)x_+\|_2 < \lambda\|x_-\|_2 + (1-\lambda)\|x_+\|_2$$

So

$$\frac{\|x\|_1}{\|x\|_2} > \frac{\lambda\|x_-\|_1 + (1-\lambda)\|x_+\|_1}{\lambda\|x_-\|_2 + (1-\lambda)\|x_+\|_2} \geq \min\{\frac{\|x_-\|_1}{\|x_-\|_2}, \frac{\|x_+\|_1}{\|x_+\|_2}\} \qquad \square$$

By the above lemma, in order for x to obtain its minimum of $\frac{\|x\|_1}{\|x\|_2}$, every nonzero entry in x should be either '$U(x)$' or '1'. Then we have the following lemma:

Lemma 3.3. *Let $U(x) = U$. Then*

$$\frac{2\sqrt{U}}{1+U}\sqrt{\|x\|_0} \leq \frac{\|x\|_1}{\|x\|_2} \leq \sqrt{\|x\|_0}$$

Proof. To estimate the lower bound, it is reasonable to assume the number of '1' in x is l and the number of 'U' is $\|x\|_0 - l$. Then

$$g(l) := \frac{\|x\|_1}{\|x\|_2} = \frac{(1-U)l + U\|x\|_0}{\sqrt{(1-U^2)l + U^2\|x\|_0}}$$

Set $g^{'}(l) = 0$, we have $l = \frac{U}{1+U}\|x\|_0$, and the inequality of lower bound follows. □

We now prove Theorem 3.2:

Proof. Suppose x_0 is the unique solution of P_0 with a sparsity of s. First of all, by Theorem 3.1, the minimizer of P_r must be in $\mathcal{F}_L$. If any other solution $x \in \mathcal{F}_L$ satisfies $U(x) > \frac{\sqrt{\|x\|_0} - \sqrt{\|x\|_0 - s}}{\sqrt{\|x\|_0} + \sqrt{\|x\|_0 - s}}$, by solving the inequality for s, we have the following:

$$\sqrt{s} < \frac{2\sqrt{U(x)}}{1+U(x)}\sqrt{\|x\|_0}.$$

By Lemma 3.3,

$$\frac{\|x_0\|_1}{\|x_0\|_2} \leq \sqrt{s} < \frac{2\sqrt{U(x)}}{1+U(x)}\sqrt{\|x\|_0} \leq \frac{\|x\|_1}{\|x\|_2}$$

Hence solving P_r will yield the sparest solution x_0. □

3.2. Exact recovery of $l_1 - l_2$

In this subsection, we show similar exact recovery results for the difference l_1 and l_2 norms.

Lemma 3.4. *Suppose $x \geq 0 \in \mathbb{R}^n$, then*

$$\frac{\|x\|_0 - 1}{2}\min_{i \in S(x)}\{x_i\} \leq \|x\|_1 - \|x\|_2 \leq (\sqrt{\|x\|_0} - 1)\|x\|_2$$

Proof. It suffices to show the lower bound given that the upper bound is directly by the Cauchy-Schwarz inequality.

$$\begin{aligned}
\|x\|_1 - \|x\|_2 &= \frac{\|x\|_1^2 - \|x\|_2^2}{\|x\|_1 + \|x\|_2} \\
&= \frac{\sum_{i \neq j}^{n} x_i x_j}{\|x\|_1 + \|x\|_2} \\
&= \frac{\sum_{i \neq j \in S(x)} x_i x_j}{\|x\|_1 + \|x\|_2} \\
&\geq \frac{(\|x\|_0 - 1)\|x\|_1 \min_{i \in S(x)}\{x_i\}}{\|x\|_1 + \|x\|_2} \\
&= \frac{\|x\|_0 - 1}{1 + \frac{\|x\|_2}{\|x\|_1}} \min_{i \in S(x)}\{x_i\} \\
&\geq \frac{\|x\|_0 - 1}{2} \min_{i \in S(x)}\{x_i\}
\end{aligned}$$ □

Theorem 3.3. *If x_0 uniquely solves P_0 with a sparsity of s, and if $\min_{i \in S(x)}\{x_i\} > \frac{2(\sqrt{s}-1)}{\|x\|_0 - 1}\|x_0\|_2$, $\forall\, x \in \mathcal{F}_L \setminus \{x_0\}$, then x_0 also uniquely solves P_d.*

Proof. Suppose $\min_{i \in S(x)}\{x_i\} > \frac{2(\sqrt{s}-1)}{\|x\|_0 - 1}\|x_0\|_2$, $\forall\, x \in \mathcal{F}_L \setminus \{x_0\}$, then by Lemma 3.4,

$$\begin{aligned}
\|x_0\|_1 - \|x_0\|_2 &\leq (\sqrt{s} - 1)\|x_0\|_2 \\
&< \frac{\|x\|_0 - 1}{2} \min_{i \in S(x)}\{x_i\} \\
&\leq \|x\|_1 - \|x\|_2, \ \forall\, x \in \mathcal{F}_L \setminus \{x_0\}
\end{aligned}$$

Moreover, the solution of P_d is contained in $\mathcal{F}_L$ by Theorem 3.1. Hence x_0 is the unique solution of P_d. □

4. Numerical approach

In this section, we consider the numerical aspects of minimizing l_1/l_2 and $l_1 - l_2$ penalties for finding sparse solutions. The setting where it is most effective computationally is in the unconstrained optimization model:

$$\min_{x \in X} F(x) := \frac{1}{2}\|A\,x - b\|^2 + R(x), \tag{4.5}$$

where $R(x) = \gamma\frac{\|x\|_1}{\|x\|_2}$ or $R(x) = \gamma(\|x\|_1 - \|x\|_2)$, and $X = \{x \in \mathbb{R}^N : x_i \geq 0, \sum_i x_i \geq r > 0\}$. Due to the nonnegativity constraint, $R(x)$ simplifies to $\gamma\frac{\langle \mathbf{1}, x\rangle}{\|x\|_2}$, where $\mathbf{1}$ denotes the constant vector in $\mathbb{R}^N$ consisting of all ones. The model (4.5) allows some measurement error in representing b in terms of the coherent dictionary, and helps to regularize the ill-conditioning of A.

Under the nonnegative constraints, it is reasonable to assume that $F(x)$ is coercive on X in the sense that for any $x^0 \in X$ the set $\{x \in X : F(x) \leq F(x^0)\}$ is bounded. This is true if there are no nonnegative vectors in $\ker(A)$, which follows for example if A has only nonnegative elements and no columns that are identically zero. Let us consider the more challenging ratio penalty first. Since R is differentiable on X, it is natural to use a gradient projection approach to solve (4.5). We will use the scaled gradient projection method proposed for a similar class of problems in [6]. The approach is based on the estimate

$$\begin{aligned} F(y) - F(x) \leq{}& (y-x)^T((\lambda_R - \frac{1}{2}\lambda_r)\mathcal{I} - C)(y-x) \\ &+ (y-x)^T(\frac{1}{2}A^TA + C)(y-x) + (y-x)^T\nabla F(x), \end{aligned}$$

where λ_r and λ_R are lower and upper bounds respectively on the eigenvalues of $\nabla^2 R(x)$ for $x \in X$ and C is any matrix. This leads naturally to the strategy of iterating

$$x^{n+1} = \arg\min_{x\in X}(x-x^n)^T(\frac{1}{2}A^TA + c_n\mathcal{I})(x-x^n) + (x-x^n)^T\nabla F(x^n). \tag{4.6}$$

To ensure convergence and a monotonically decreasing objective $F(x^n)$, it suffices to choose $c_n > 0$ such that there is a sufficient decrease in F according to

$$\begin{aligned} F(x^{n+1}) - F(x^n) \leq \sigma[&(x^{n+1}-x^n)^T(\frac{1}{2}A^TA + c_n\mathcal{I})(x^{n+1}-x^n) \\ &+ (x^{n+1}-x^n)^T\nabla F(x^n)] \end{aligned} \tag{4.7}$$

for some $\sigma \in (0, 1]$. To improve the method's overall efficiency, c_n can be adjusted every iteration to prefer smaller values while still ensuring a sufficient decrease in F. The complete algorithm from [6] is shown below for reader's convenience.

Algorithm 1. A Scaled Gradient Projection Method for Solving (4.5) with $R(x) = \gamma\frac{\|x\|_1}{\|x\|_2}$.

Define $x^0 \in X$, $c_0 > 0$, $\sigma \in (0,1]$, $\epsilon_1 > 0$, $\rho > 0$, $\xi_1 > 1, \xi_2 > 1$ and set $n = 0$.

`while` $n = 0$ or $\|x^n - x^{n-1}\|_\infty > \epsilon_1$

$$y = \arg\min_{x \in X} (x - x^n)^T(\frac{1}{2}A^TA + c_n\mathcal{I})(x - x^n) + (x - x^n)^T\nabla F(x^n)$$

`if` $F(y) - F(x^n) > \sigma\left[(y - x^n)^T(\frac{1}{2}A^TA + c_n\mathcal{I})(y - x^n) + (y - x^n)^T\nabla F(x^n)\right]$

$c_n = \xi_2 c_n$

`else`

$x^{n+1} = y$

$$c_{n+1} = \begin{cases} \frac{c_n}{\xi_1} & \text{if smallest eigenvalue of } \frac{c_n}{\xi_1}\mathcal{I} + \frac{1}{2}A^TA \text{ is greater than } \rho \\ c_n & \text{otherwise.} \end{cases}$$

$n = n + 1$

`end if`

`end while`

Any limit point x^* of the sequence of iterates $\{x^n\}$ satisfies $(y - x^*)^T\nabla F(x^*) \geq 0$ for all $y \in X$ and is therefore a stationary point of (4.5) [6]. Note that every iteration requires solving the convex problem

$$\min_{x \in X} (x - x^n)^T(\frac{1}{2}A^TA + c_n\mathcal{I})(x - x^n) + (x - x^n)^T\nabla F(x^n).$$

As in [6] we can solve this using the Alternating Direction Method of Multipliers (ADMM) [8, 9]. The explicit iterations are described in the following algorithm.

Algorithm 2. ADMM for solving convex subproblem.

Define $\delta > 0$, $\epsilon_2 > 0$, v^0 and p^0 arbitrarily and let $k = 0$.

`while` $k = 0$ or $\|\frac{\|v^k - v^{k-1}\|}{\|v^k - x^n\|} > \epsilon_2$ or $\frac{\|v^k - u^k\|}{\|v^k - x^n\|} > \epsilon_2$

$$u^{k+1} = x^n + (A^TA + (2c_n + \delta)\mathcal{I})^{-1}\left(\delta(v^k - x^n) - p^k - \nabla F(x^n)\right)$$

$$v^{k+1} = \Pi_X\left(u^{k+1} + \frac{p^k}{\delta}\right)$$

$$p^{k+1} = p^k + \delta(u^{k+1} - v^{k+1})$$

$k = k + 1$

`end while`

$x^{n+1} = v^k.$

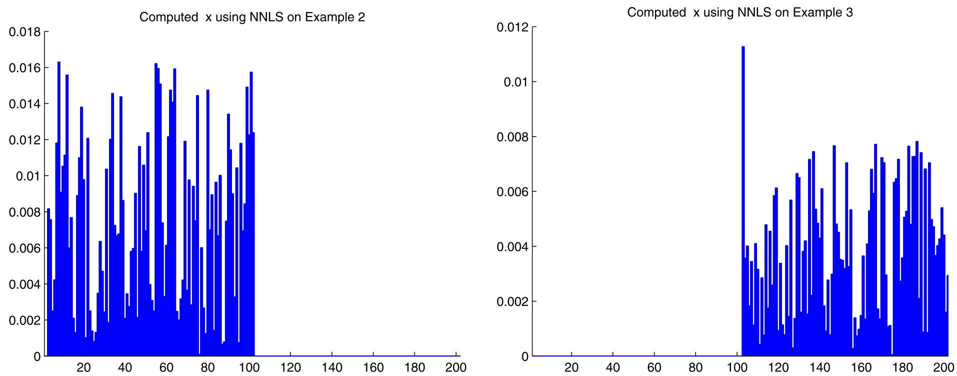

Figure 2: Estimated x using non-negative least squares.

Here, Π_X denotes the orthogonal projection onto X. Note that if AA^T is much smaller in size than A^TA we can use the Woodbury identity to rewrite the inverse that appears in Algorithm 2.

As numerical experiments we apply these algorithms to Examples 2 and 3, with both of the A matrices defined using values of $n = 100$, $p = 0.95$ and b a vector of n random numbers uniformly distributed on $[0, 1]$. The coherence and ill-conditioning of these matrices make these examples numerically challenging. Non-negative least squares, often a good method for finding sparse nonnegative solutions when they exist [1], fails to find sparse solutions for these examples as shown in Figure 2. Solving the $\frac{l_1}{l_2}$ model (4.5) on the other hand, while it does not identify the sparsest solutions, does find solutions with much better sparsity properties. The results for Examples 2 and 3 are shown in Figure 3. The model parameters used were $\gamma = 0.1$ and $r = 0.05$. For the algorithm parameters, $\delta = 1$, $c_0 = 10^{-9}$, $\xi_1 = 2$, $\xi_2 = 10$ and $\sigma = 0.01$. The most important of the algorithm parameters is δ, which affects the efficiency of ADMM on the convex subproblem. The tolerances for the stopping conditions were set to $\epsilon_1 = 10^{-8}$ and $\epsilon_2 = 10^{-4}$.

For Example 2, Algorithm 1 recovered the 2-sparse solution $[1, 1, 0, \cdots, 0]'$. For Example 3 it approximately recovered $[0, 0, (e^1)', \frac{(b^1)'}{a}]'$, which has the property that one coefficient is much larger than all the others.

These results are initialization dependent. Here we initialized x^0 to be a constant vector, which is partly to blame for finding a 2-sparse solution to Example 2 that is a stationary point but not a local minimum. Instead, consider initializing x^0 to be a small perturbation of a constant vector, for instance $x^0_i = r(100 + 0.01\eta_i)$ with η_i sampled from a normal distribution with mean zero and standard deviation 1. With such an initialization, we

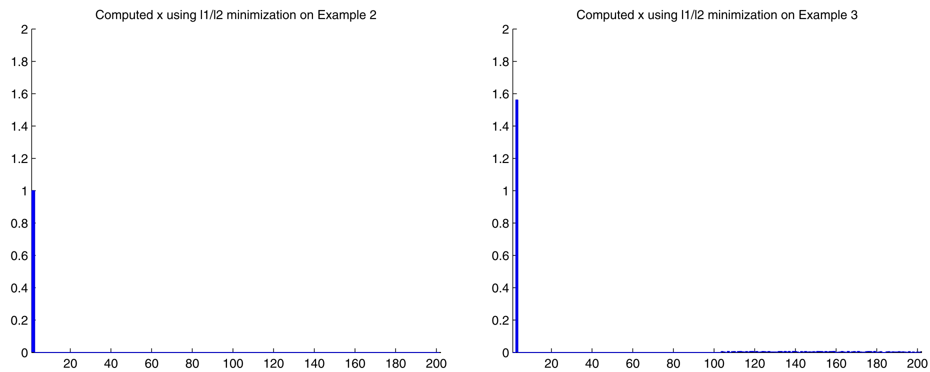

Figure 3: Estimated x using Algorithm 1 on (4.5).

are far more likely to find one of the 1-sparse solutions $[2,0,0,\cdots,0]'$ or $[0,2,0,\cdots,0]'$.

Another important numerical consideration is the parameter r that acts as a lower bound on the l_1 norms of the possible solutions. Because of the way the matrices A are scaled for Examples 2 and 3, the sparsest solutions also have larger l_1 norms. In this case, larger values of r promote sparsity. Choosing $r = 0.05$ is still much less than the norms of the NNLS solutions, so it is not the case here that those potential solutions were eliminated by the choice of constraint set.

Minimizing the difference of l_1 and l_2 norms is easier than minimizing the ratio because the objective becomes a difference of convex functions. In particular we can set $c_n = c$ for any $c > 0$ in the iteration (4.6) and be guaranteed to satisfy the sufficient decrease inequality (4.7) with $\sigma = 1$. Moreover, since the difference penalty is better behaved at the origin, we could consider simplifying the constraint set X and letting it be the entire nonnegative orthant. However, we choose to leave X as previously defined since it may be advantageous to disallow solutions whose l_1 norms are below some threshold r. Using a constant c, Algorithm 1 can be simplified to the following.

Algorithm 3. SGP Method for Solving (4.5) with $R(x) = \gamma(||x||_1 - ||x||_2)$.

Define $x^0 \in X$, $c > 0$, $\epsilon_1 > 0$ and set $n = 0$.

`while` $n = 0$ or $\|x^n - x^{n-1}\|_\infty > \epsilon_1$

$$x^{n+1} = \arg\min_{x\in X} (x - x^n)^T(\frac{1}{2}A^T A + c\mathcal{I})(x - x^n) + (x - x^n)^T \nabla F(x^n)$$

$n = n + 1$

`end while`

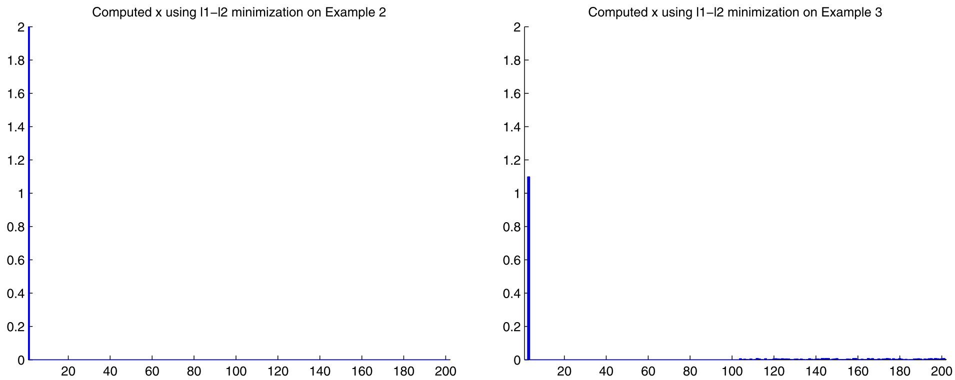

Figure 4: Estimated x using Algorithm 3 on (4.5).

Algorithm 2 can again be used to solve the convex subproblem in Algorithm 3.

We repeat the experiments on Examples 2 and 3 using Algorithm 3 to numerically compare how well the $l_1 - l_2$ penalty is able to promote sparsity. We first attempt to use the same parameters as before, setting $\gamma = 0.1$, $r = 0.05$, $\delta = 1$ and $c = 10^{-9}$. We again set the tolerances for the stopping conditions to be $\epsilon_1 = 10^{-8}$ for the outer iterations and $\epsilon_2 = 10^{-4}$ for the inner iterations. Unfortunately, with these parameters $l_1 - l_2$ minimization does not yield sparse solutions for either Example 2 or 3. Two approaches to improve sparsity are to increase γ or to increase r. Using large values of γ does yield sparse vectors, but they are highly sensitive to the initialization and are often not close to the correct sparse solutions. On the other hand, if we keep all the parameters the same but increase r to $r = 0.5$, then we are able to get the sparse solutions shown in Figure 4, which are similar to those generated by Algorithm 1. For Example 2, the l_1 norm of the NNLS solution is approximately 0.76, so it is still in our constraint set. For Example 3, however, the l_1 norm of the NNLS solution is approximately 0.39, which falls outside our constraint set when we set $r = 0.5$. So the sparse result for Example 3 shown in Figure 4 is special to this problem and probably has more to do with the constraint set than it does with $l_1 - l_2$ minimization. But for Example 2, $l_1 - l_2$ minimization did help find a good sparse solution.

5. Discussion and conclusion

We studied properties of the ratio and difference of l_1 and l_2 norms in finding sparse solutions from a representation with coherent and redundant dictionaries. We presented an exact recovery theory and showed both anlaytical and numerical examples. In future work, we plan to investigate further the mathematical theory and computational performance of the related algorithms based on these sparsity promoting measures, also apply them to data in applications. A work along this line is [6].

Acknowledgments

The work was partially supported by NSF grants DMS-0911277, DMS-0928427, and DMS-1222507. The authors would like to thank Professors Russel Caflisch, Ingrid Daubechies, Tom Hou, and Stanley Osher for their interest, and the opportunity to present some of the results here at the Adaptive Data Analysis and Sparsity Workshop at the Institute for Pure and Applied Mathematics at UCLA, Jan. 31, 2013.

References

[1] A. Bruckstein, M. Elad, and M. Zibulevsky, On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations, *IEEE Transactions on Information Theory*, 54(11), 4813–4820, Nov. 2008. MR2589866

[2] E. Candès, J. Romberg, and T. Tao, Robust uncertainty principles: Exact signal rconstruction from highly incomplete Fourier information, *IEEE Trans. Info. Theory*, 52(2), 489–509, Feb. 2006. MR2236170

[3] E. Candès and T. Tao, Near optimal signal recovery from random projections: Universal encoding stratgies?, *IEEE Trans. Info. Theory*, 52(12), 5406–5425, Dec. 2006. MR2300700

[4] D. Donoho, Compressed sensing, *IEEE Trans. Info. Theory*, 52(4), 1289–1306, April, 2006. MR2241189

[5] D. Donoho and M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization, *Proc. Nat. Acad. Scien. USA*, 100, 2197–2202, Mar. 2003. MR1963681

[6] E. Esser, Y. Lou, and J. Xin, A method for finding structured sparse solutions to non-negative least squares problems with applications, *SIAM J. Imaging Sciences*, 6(4), 2010–2046, 2013. MR3118253

[7] A. Fannjiang and W. Liao, Coherence pattern-guided compressive sensing with unresolved grids, *SIAM J. Imaging Sciences*, 5(1), 179–202, 2012. MR2902661

[8] D. Gabay and B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite-element approximations, *Comp. Math. Appl.*, 2, 17–40, 1976.

[9] R. Glowinski and A. Marrocco, Sur lapproximation par elements finis dordre un, et la resolution par penalisation-dualite dune classe de problemes de Dirichlet nonlineaires, *Rev. Francaise dAut Inf. Rech. Oper.*, R-2, 41–76, 1975. MR0388811

[10] W. Hartmann, "Signals, Sound, and Sensation", AIP Press, 1998, Chapter 10 (Auditory Filters).

[11] P. Hoyer, Non-negative matrix factorization with sparseness constraints, *Journal of Machine Learning Research*, 5(12), 1457–1469, 2004. MR2248024

[12] D. Krishnan, T. Tay, and R. Fergus, Blind deconvolution using a normalized sparsity measure, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[13] H. Ji, J. Li, Z. Shen, and K. Wang, Image deconvolution using a characterization of sharp images in wavelet domain, *Applied and Computational Harmonic Analysis*, 32(2), 295–304, 2012. MR2880285

[14] B. Olshaussen and D. Field, Sparse coding with an overcomplete basis set: A strategy employed by V1?, *Vision Research*, 37, 3311–3325, 1997.

[15] P. D. Tao and L. T. H. An, Convex analysis approach to d.c. programming: Theory, algorithms and applications, *Acta Mathematica Vietnamica*, 22(1), 289–355, 1997. MR1479751

[16] J. Xin and Y.-Y. Qi, "Mathematical Modeling and Signal Processing in Speech and Hearing Sciences", MS&A, Vol. 10, Springer, 2014, Chapter 4 (Speech Recognition). MR3184779

Penghang Yin
Department of Mathematics
University of California
Irvine, CA, 92697
USA
E-mail address: penghany@uci.edu

Ernie Esser
Department of Mathematics
University of California
Irvine, CA, 92697
USA
E-mail address: eesser@math.uci.edu

Jack Xin
Department of Mathematics
University of California
Irvine, CA, 92697
USA
E-mail address: jxin@math.uci.edu

Received May 23, 2014