FAST LASSO SCREENING TESTS BASED ON CORRELATIONS

Zhen James Xiang and Peter J. Ramadge

Dept. of Electrical Engineering, Princeton University, Princeton NJ

ABSTRACT

Representing a vector as a sparse linear combination of codewords, e.g. by solving a lasso problem, lies at the heart of many machine learning and statistics applications. To improve the efficiency of solving lasso problems, we systematically investigate lasso screening, a process that quickly identifies dictionary entries that won't be used in the optimal sparse representation, and hence can be removed from the problem. We propose a general test called an $\mathcal{R}$ region test that unifies existing screening tests and we derive a particular instance called the dome test. This test is stronger than existing screening tests and can be executed in linear-time as a two-pass test with a memory footprint of only three codewords.

Index Terms— Machine learning, Optimization, Algorithms.

1. INTRODUCTION

For $\mathbf{x}, \mathbf{b}_i \in \mathbb{R}^p$, $i = 1, \ldots, m$, consider the lasso problem:

$$\min_{w_1,w_2,\ldots,w_m} \quad \tfrac{1}{2}\|\mathbf{x} - \textstyle\sum_{i=1}^m w_i \mathbf{b}_i\|_2^2 + \lambda \sum_{i=1}^m |w_i|, \ (1)$$

where $\mathbf{x}$ and $\mathbf{b}_i$ are assumed to be normalized: $\|\mathbf{x}\|_2 = \|\mathbf{b}_i\|_2 = 1$, $i = 1, \ldots, m$. This problem arises frequently in machine learning and statistics applications. The idea is to "encode" $\mathbf{x}$ as a sparse combination of the "codewords" or "atoms" in the dictionary $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m]$. The weights $\tilde{w}_i$ obtained by solving (1) are then used for subsequent processing (e.g. classification). Such sparse representation methods have exhibited superior performance on difficult computer vision problems such as face [1] and object [2] recognition. In these applications $\mathbf{x}$ is usually highly correlated with some of the codewords, e.g., a codeword in the same class as $\mathbf{x}$.

The challenge is that solving (1) becomes harder when m is large. One way to address this is by screening the dictionary: If we can determine via a simple test that $\tilde{w}_i = 0$, then $\mathbf{b}_i$ can be excluded from the actual optimization. This effectively reduces m in problem (1). Such tests should be fast, e.g. linear-time complexity $O(mp)$. As examples, the SAFE rule in [3] rejects $\mathbf{b}_i$ when $|\mathbf{x}^T\mathbf{b}_i| < \lambda + \lambda/\lambda_{\max} - 1$ where $\lambda_{\max} = \max_i |\mathbf{x}^T\mathbf{b}_i|$, and the basic strong rule in [4] rejects $\mathbf{b}_i$ when $|\mathbf{x}^T\mathbf{b}_i| < 2\lambda - \lambda_{\max}$ (although it should be noted that this rule can falsely reject some $\mathbf{b}_i$). A general "sphere test" is proposed in [5], in which the SAFE rule is characterized as a specific sphere test and a useful alternative sphere test ST2, which thresholds $|\mathbf{x}^T\mathbf{b}_i|$ at a different level, is introduced. The strongest test in [5] (ST3) uses a second correlation between $\mathbf{b}_i$ and $\mathbf{b}_*$, a vector in $\{\pm\mathbf{b}_i\}_{i=1}^m$ with $\mathbf{x}^T\mathbf{b}_* = \lambda_{\max}$. This test rejects $\mathbf{b}_i$ by thresholding a linear combination of the correlations $\mathbf{x}^T\mathbf{b}_i$ and $\mathbf{b}_*^T\mathbf{b}_i$.

The above tests are all based on bounding the optimal solution $\tilde{\boldsymbol{\theta}}$ of the dual problem of (1) within a sphere (hence the name "sphere test"). In this paper we introduce a more general concept called an $\mathcal{R}$ region test, which is based on bounding $\tilde{\boldsymbol{\theta}}$ within a region $\mathcal{R}$. As $\mathcal{R}$ shrinks, the test becomes stronger but more time consuming. To achieve a good tradeoff, we use two simple constraints to bound $\tilde{\boldsymbol{\theta}}$ within a spherical dome region $\mathcal{G}$ and derive a new test called the dome test. All previous tests are based on spheres containing $\mathcal{G}$ and are all weaker than the new test. In addition, we specialize the test to a screening test based only on $\mathbf{x}^T\mathbf{b}_i$. This single correlation test is stronger than all previous tests in the literature that only use the correlations $\mathbf{x}^T\mathbf{b}_i$.

The paper is organized as follows. We introduce the $\mathcal{R}$ region test in §2 and setup the framework of the dome test. We outline the derivation of the dome test in §3, empirically evaluate its performance in §4 and conclude in §5.

2. PRELIMINARIES

We begin by considering the dual problem of (1):

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{\lambda^2}{2}\|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda}\|_2^2 \\ \text{s.t.} \quad & |\boldsymbol{\theta}^T\mathbf{b}_i| \le 1 \quad \forall i = 1, 2, \ldots, m. \end{aligned} \qquad (2)$$

The solutions $\tilde{\mathbf{w}} = (\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_m)^T$ and $\tilde{\boldsymbol{\theta}}$ of the primal and dual problems are related through (see [5]):

$$\mathbf{x} = \sum_{i=1}^m \tilde{w}_i \mathbf{b}_i + \lambda\tilde{\boldsymbol{\theta}}, \ \tilde{\boldsymbol{\theta}}^T\mathbf{b}_i \in \left\{ \begin{array}{rl} \{\text{sign}\, \tilde{w}_i\} & \text{if } \tilde{w}_i \neq 0, \\ [-1, 1] & \text{if } \tilde{w}_i = 0. \end{array} \right. \quad (3)$$

By (3), if $|\tilde{\boldsymbol{\theta}}^T\mathbf{b}_i| < 1$, then $\tilde{w}_i = 0$. Hence the dual problem provides a sufficient condition for excluding $\mathbf{b}_i$:

$$|\tilde{\boldsymbol{\theta}}^T\mathbf{b}_i| < 1 \Rightarrow \tilde{w}_i = 0. \qquad (4)$$

This research partially supported by NSF grant CCF-1116208.

This key observation is the basis for all that follows. In light of its importance we call (4) the *core rejection test* and denote it by $T_{\{\tilde{\boldsymbol{\theta}}\}}$. The test is not practical since its application requires knowledge of $\tilde{\boldsymbol{\theta}}$, the solution of the entire lasso dual problem. A practical way to use (4) is to extract partial information that bounds $\tilde{\boldsymbol{\theta}}$ in a region $\mathcal{R} \subset \mathbb{R}^p$. Given such $\mathcal{R}$, we define the $\mathcal{R}$ region test $T_{\mathcal{R}}$ as:

$$(-1 < \min_{\boldsymbol{\theta}\in\mathcal{R}} \boldsymbol{\theta}^T\mathbf{b}_i \text{ and } \max_{\boldsymbol{\theta}\in\mathcal{R}} \boldsymbol{\theta}^T\mathbf{b}_i < 1) \Rightarrow \tilde{w}_i = 0. \tag{5}$$

When $\mathcal{R} = \{\tilde{\boldsymbol{\theta}}\}$, (5) reduces to (4).

A test T' complies with test T if T' only rejects codewords that T also rejects. We only consider tests that comply with the core rejection test $T_{\{\tilde{\boldsymbol{\theta}}\}}$. It's easy to verify that if $\mathcal{R}_2 \subset \mathcal{R}_1$, then $T_{\mathcal{R}_1}$ complies with $T_{\mathcal{R}_2}$ Therefore, if $\tilde{\boldsymbol{\theta}}$ is bounded by $\mathcal{R}$, i.e., $\{\tilde{\boldsymbol{\theta}}\} \subset \mathcal{R}$, then $T_{\mathcal{R}}$ complies with the core rejection test. A tighter $\mathcal{R}$ yields a stronger test. However, there is a tradeoff between how tightly $\mathcal{R}$ bounds $\tilde{\boldsymbol{\theta}}$ and the time complexity of the test. For example, the core rejection test $T_{\{\tilde{\boldsymbol{\theta}}\}}$ is very strong but requires solving the entire dual lasso problem. At the other extreme, choosing $\mathcal{R} = \mathbb{R}^p$ yields no rejection but calculating $\min_{\boldsymbol{\theta}\in\mathcal{R}} \boldsymbol{\theta}^T\mathbf{b}_i$ and $\max_{\boldsymbol{\theta}\in\mathcal{R}} \boldsymbol{\theta}^T\mathbf{b}_i$ is trivial. To achieve a good tradeoff, we seek a tight bounding region $\mathcal{R}$ such that $T_{\mathcal{R}}$ can be efficiently executed, e.g. linear time complexity.

Our approach to selecting a suitable region exploits the observation made earlier that in many relevant applications there is at least one codeword that has high absolute correlation with $\mathbf{x}$. Let $\mathbf{b}_*$ be selected from the set $\{\pm\mathbf{b}_i\}_{i=1}^m$ so as to maximize $\mathbf{x}^T\mathbf{b}_*$ and set $\lambda_{\max} = \mathbf{x}^T\mathbf{b}_*$. Note that $\mathbf{x}/\lambda_{\max}$ is a feasible solution of (2) since $\forall i : |\mathbf{x}^T\mathbf{b}_i/\lambda_{\max}| \le 1$. By (2), $\tilde{\boldsymbol{\theta}}$ is the closest feasible point to $\mathbf{x}/\lambda$. Therefore $\tilde{\boldsymbol{\theta}}$ is contained in the closed ball:

$$\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \mathbf{x}/\lambda\|_2 \le \|\mathbf{x}/\lambda_{\max} - \mathbf{x}/\lambda\|_2 = 1/\lambda - 1/\lambda_{\max}\}. \tag{6}$$

Intersecting this ball with the constraint $\boldsymbol{\theta}^T\mathbf{b}_* \le 1$ further bounds $\tilde{\boldsymbol{\theta}}$ within the spherical dome region:

$$\mathcal{G} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \mathbf{x}/\lambda\|_2 \le 1/\lambda - 1/\lambda_{\max}, \mathbf{b}_*^T\boldsymbol{\theta} \le 1\}. \tag{7}$$

We call $\mathcal{G}$ the $\mathbf{b}_*$-dome and $T_{\mathcal{G}}$ the $\mathbf{b}_*$-dome test (or simply *dome test*). For the rest of the paper we examine the test $T_{\mathcal{G}}$.

Before proceeding, we relate the dome test to other tests in the literature. The tests ST1 (the SAFE rule of [3]), ST2 and ST3 in [5] are derived via relaxations of $\mathcal{G}$ to spheres. Hence these tests all comply with the dome test $T_{\mathcal{G}}$ and therefore are all weaker than $T_{\mathcal{G}}$. The basic strong rule of [4] doesn't comply with the core rejection test and is not considered further.

3. THE DOME TEST

We now unfold (5) with $\mathcal{R} = \mathcal{G}$ into a simple test on the correlations $\mathbf{x}^T\mathbf{b}_i$ and $\mathbf{b}_*^T\mathbf{b}_i$, $i = 1, \dots, m$. To proceed, let

$$\underline{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}\in\mathcal{G}} \boldsymbol{\theta}^T\mathbf{b}_i \quad \text{and} \quad \bar{\boldsymbol{\theta}}_i = \arg\max_{\boldsymbol{\theta}\in\mathcal{G}} \boldsymbol{\theta}^T\mathbf{b}_i. \tag{8}$$

The test $T_{\mathcal{G}}$ checks the conditions $-1 < \underline{\boldsymbol{\theta}}_i^T\mathbf{b}_i$ and $\bar{\boldsymbol{\theta}}_i^T\mathbf{b}_i < 1$. The expressions for $\underline{\boldsymbol{\theta}}_i^T\mathbf{b}_i$ and $\bar{\boldsymbol{\theta}}_i^T\mathbf{b}_i$ depend on the locations of $\underline{\boldsymbol{\theta}}_i$ and $\bar{\boldsymbol{\theta}}_i$. $\bar{\boldsymbol{\theta}}_i$ and $\underline{\boldsymbol{\theta}}_i$ must lie on the boundary of $\mathcal{G}$ (because $\boldsymbol{\theta}^T\mathbf{b}_i$ is linear in $\boldsymbol{\theta}$ and $\mathcal{G}$ is closed). $\mathcal{G}$ is enclosed by two boundaries: a spherical cap boundary $\mathcal{C}$ with $\mathbf{b}_*^T\boldsymbol{\theta} < 1$ and equality in (6), and a disc boundary $\mathcal{D}$ consisting of the the subset of the hyperplane $\mathbf{b}_*^T\boldsymbol{\theta} = 1$ inside the ball (6). It's easy to work out from the geometry that the disc has center $\mathbf{q} = \mathbf{x}/\lambda - (\lambda_{\max}/\lambda - 1)\mathbf{b}_*$ and radius $r = \sqrt{1/\lambda_{\max}^2 - 1}(\lambda_{\max}/\lambda - 1)$. The following lemma indicates which boundary contains $\underline{\boldsymbol{\theta}}_i$ and $\bar{\boldsymbol{\theta}}_i$.

Lemma 1. *$\bar{\boldsymbol{\theta}}_i \in \mathcal{C}$ if and only if $\mathbf{b}_*^T\mathbf{b}_i < -\lambda_{\max}$; and $\underline{\boldsymbol{\theta}}_i \in \mathcal{C}$ if and only if $\mathbf{b}_*^T\mathbf{b}_i > \lambda_{\max}$.*

Lemma 1 indicates that we must consider three cases:
(i): $\mathbf{b}_*^T\mathbf{b}_i < -\lambda_{\max}$. In this case $\bar{\boldsymbol{\theta}}_i \in \mathcal{C}$ and $\underline{\boldsymbol{\theta}}_i \in \mathcal{D}$.
(ii): $\mathbf{b}_*^T\mathbf{b}_i > \lambda_{\max}$. In this case $\bar{\boldsymbol{\theta}}_i \in \mathcal{D}$ and $\underline{\boldsymbol{\theta}}_i \in \mathcal{C}$.
(iii): $-\lambda_{\max} \le \mathbf{b}_*^T\mathbf{b}_i \le \lambda_{\max}$. In this case $\bar{\boldsymbol{\theta}}_i, \underline{\boldsymbol{\theta}}_i \in \mathcal{D}$.

When $\underline{\boldsymbol{\theta}}_i$ (resp. $\bar{\boldsymbol{\theta}}_i$) is known to be in a given boundary ($\mathcal{D}$ or $\mathcal{C}$), we can readily obtain closed form expressions for $\underline{\boldsymbol{\theta}}_i$ and $\underline{\boldsymbol{\theta}}_i^T\mathbf{b}_i$ (resp. $\bar{\boldsymbol{\theta}}_i$ and $\bar{\boldsymbol{\theta}}_i^T\mathbf{b}_i$). These results are summarized in the following lemma:

Lemma 2.

$$\bar{\boldsymbol{\theta}}_i^T\mathbf{b}_i = \begin{cases} \mathbf{q}^T\mathbf{b}_i + r\|(I - \mathbf{b}_*\mathbf{b}_*^T)\mathbf{b}_i\|_2, & \textit{if } \bar{\boldsymbol{\theta}}_i \in \mathcal{D}; \\ \mathbf{x}^T\mathbf{b}_i/\lambda + (1/\lambda - 1/\lambda_{\max}), & \textit{if } \bar{\boldsymbol{\theta}}_i \in \mathcal{C}. \end{cases} \tag{9}$$

$$\underline{\boldsymbol{\theta}}_i^T\mathbf{b}_i = \begin{cases} \mathbf{q}^T\mathbf{b}_i - r\|(I - \mathbf{b}_*\mathbf{b}_*^T)\mathbf{b}_i\|_2, & \textit{if } \underline{\boldsymbol{\theta}}_i \in \mathcal{D}; \\ \mathbf{x}^T\mathbf{b}_i/\lambda - (1/\lambda - 1/\lambda_{\max}), & \textit{if } \underline{\boldsymbol{\theta}}_i \in \mathcal{C}. \end{cases} \tag{10}$$

Lemmas 1 and 2 are not difficult but the proofs (see [6]) are lengthy. Using these two lemmas, the $T_{\mathcal{G}}$ test can be obtained by substituting the correct equation from (9) and (10) into $-1 < \underline{\boldsymbol{\theta}}_i^T\mathbf{b}_i$ and $\bar{\boldsymbol{\theta}}_i^T\mathbf{b}_i < 1$ for the three difference cases (i)-(iii). This yields the following result:

Theorem 1. *The dome test $T_{\mathcal{G}}$ is characterized by*

$$\big(Q_l(\mathbf{b}_*^T\mathbf{b}_i) < \mathbf{x}^T\mathbf{b}_i < Q_u(\mathbf{b}_*^T\mathbf{b}_i)\big) \Rightarrow \tilde{w}_i = 0. \tag{11}$$

Where $Q_l(t)$ and $Q_u(t)$ are C^1 functions defined by:

$$Q_l(t) = \begin{cases} (\lambda_{\max} - \lambda)t - \lambda + \lambda r\sqrt{1 - t^2}, & t \le \lambda_{\max}; \\ -(\lambda - 1 + \lambda/\lambda_{\max}), & t > \lambda_{\max}. \end{cases}$$

$$Q_u(t) = \begin{cases} (\lambda - 1 + \lambda/\lambda_{\max}), & t < -\lambda_{\max}; \\ (\lambda_{\max} - \lambda)t + \lambda - \lambda r\sqrt{1 - t^2}, & t \ge -\lambda_{\max}. \end{cases}$$

Notice that (11) depends only on the correlations $\mathbf{x}^T\mathbf{b}_i$ and $\mathbf{b}_*^T\mathbf{b}_i$. Hence the dome test can be executed using only the $2m$ correlations $\{\mathbf{x}^T\mathbf{b}_i, \mathbf{b}_*^T\mathbf{b}_i\}_{i=1}^m$. Finding $\mathbf{b}_*$ and $\lambda_{\max}$ in the first place requires calculating m correlations. Since each of these correlations can be computed in $O(p)$ time, the dome test has $O(mp)$ time complexity with a very small constant. It remains to determine how successfully the test rejects codewords and what improvement it offers over existing tests. This is examined empirically in §4.

3.1. The dome test restricted to $\mathbf{x}^T\mathbf{b}_i$

The dome test and the test ST3 use two correlations: $\mathbf{x}^T\mathbf{b}_i$ and $\mathbf{b}_*^T\mathbf{b}_i$. In contrast, simpler tests threshold only $|\mathbf{x}^T\mathbf{b}_i|$: $|\mathbf{x}^T\mathbf{b}_i| < v \Rightarrow \tilde{w}_i = 0$ (see [3],[5]). Such tests have a simple intuition: codewords less correlated with $\mathbf{x}$ are discarded. There is a natural way to restrict the dome test to the correlation $\mathbf{x}^T\mathbf{b}_i$. This restriction is

$$|\mathbf{x}^T\mathbf{b}_i| < v_* \Rightarrow \tilde{w}_i = 0, \tag{12}$$

with $v_* \geq 0$ selected to be the largest value such that (12) complies with the dome test. We call this the *restricted dome test*, or *r-dome test*. A closed form expression for v_* is derived below. Of course, r-dome is weaker than the dome test. But it's stronger than the SAFE rule in [3] and ST2 in [5].

The derivation of v_* is not difficult but is somewhat tedious. We outline the steps. For any $\mathbf{b}_i$, the projected pair $(\mathbf{b}_*^T\mathbf{b}_i, \mathbf{x}^T\mathbf{b}_i)$ lies within an ellipse with boundary:

$$\mathbf{x}^T\mathbf{b}_i = \lambda_{\max}(\mathbf{b}_*^T\mathbf{b}_i) \pm \sqrt{1-\lambda_{\max}^2}\sqrt{1-(\mathbf{b}_*^T\mathbf{b}_i)^2} \tag{13}$$

This is illustrated in Figure 2. The curve Q_u and upper ellipse boundary ((13) with + sign) are monotonic and v_* is the first (lower) intersection of Q_u and the upper edge of the ellipse boundary. Q_u has two segments indicated by the vertical dashed line. If the intersection (t_0, v_0) is on the second (non-constant) segment, then the solution is:

$$\begin{aligned} t_0 &= \frac{\lambda^2 - (\lambda r + \sqrt{1-\lambda_{\max}^2})^2}{\lambda^2 + (\lambda r + \sqrt{1-\lambda_{\max}^2})^2}, \\ v_0 &= \lambda_{\max} t_0 + \sqrt{1-\lambda_{\max}^2}\sqrt{1-t_0^2}. \end{aligned} \tag{14}$$

So if $t_0 \geq -\lambda_{\max}$, then $v_* = v_0$ in (14). Otherwise, the intersection is on the first (constant) segment of Q_u, which gives $v_* = \lambda - 1 + \lambda/\lambda_{\max}$. To summarize:

$$v_* = \begin{cases} v_0 & \text{if } t_0 \geq -\lambda_{\max}; \\ \lambda - 1 + \lambda/\lambda_{\max} & \text{otherwise.} \end{cases} \tag{15}$$

One possible application of the r-dome test is to improve efficiency: to test $\mathbf{b}_i$, first apply the r-dome test (which is faster than ST3 and dome test because only one correlation is used) and only apply the dome test if r-dome does not reject $\mathbf{b}_i$.

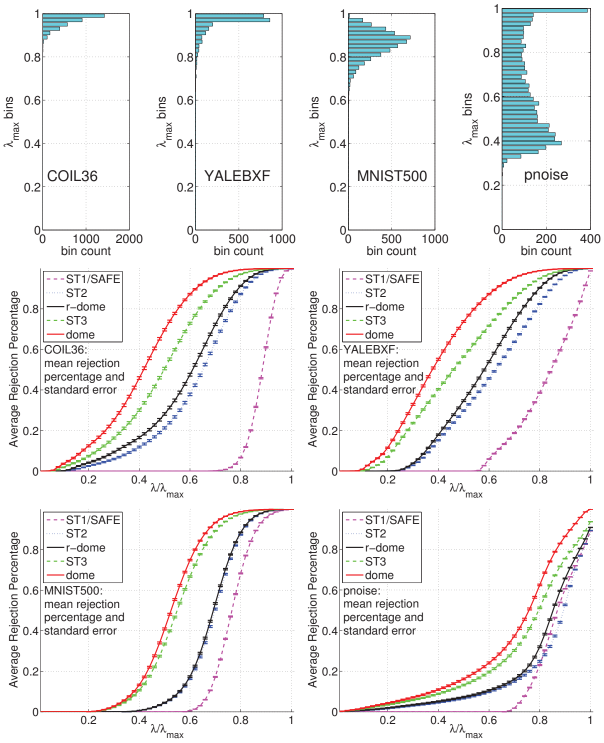

Fig. 1. Results for four data sets. For each data set: top row shows the distribution of $\lambda_{\max}$ ($\lambda_{\max}$ is consistently high for natural data sets); bottom rows show the average rejection percentage of different screening methods. The dome test outperforms all existing tests.

4. EXPERIMENTS

We now use real and synthetic data sets to compare the rejection power of various tests. We do so by recording the percentage of codewords that a test rejects, and the computation time needed to execute the test *and* solve the reduced lasso problem. We use the data sets: COIL36 ($n = 3,600$ images of size $p = 3\times128\times128 = 49,152$, we used 36 images per object in the COIL data set), YALEBXF (Yale B Extended Frontal, $n = 2,414$ frontal faces images of size $p = 192\times168 = 32,256$ of 38 subjects), and MNIST500 ($n = 5,000$ images of size $p = 28\times28 = 784$, obtained from the first 500 images of each digit in the MNIST data set). See [5] for references to these data sets. We also use a synthetic data set called "pnoise", containing $n = 5,000$ vectors of the form $\mathbf{e}_1 + 0.1\alpha\mathbf{g}$, where $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$ is the first 784-dimensional natural basis vector, α is a random variable uniformly distributed in $[0, 1]$ and $\mathbf{g}$ has distribution $\mathcal{N}(\mathbf{0}, I_{784})$ (α and $\mathbf{g}$ are independent).

All data vectors are first normalized to S^{p-1}. For a data set of n vectors, we consider n distinct lasso problems. The j-th lasso problem uses the j-th vector as $\mathbf{x}$ and the remaining vectors as codewords. For each lasso problem, we record the $\lambda_{\max}$ value and the number of rejected codewords for each test under various λ values. For each dataset we create bins of width 0.02 in $\lambda_{\max}$ and plot a histogram of the number of lasso problems falling in each bin (Figure 1). We see that for real world data sets, $\lambda_{\max}$ is usually high because the dictionary contains codewords highly correlated with $\mathbf{x}$. The "pnoise" data set is specifically designed to have a wider spread in $\lambda_{\max}$. Figure 1 also shows the mean percentage of codewords rejected averaged over the n lasso problems as $\lambda/\lambda_{\max}$ varies. As expected the dome test always has the best rejection. The improvement over ST3 is more than 10% in the

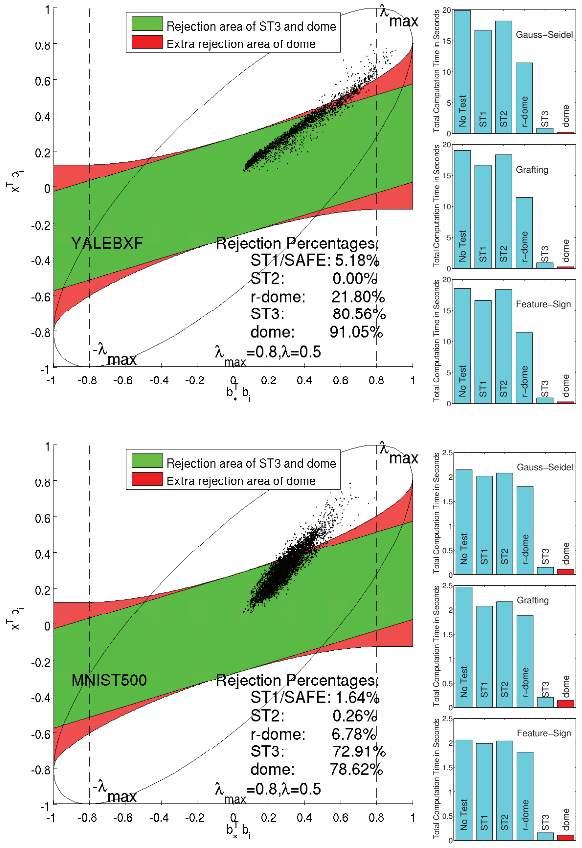

Fig. 2. For each of the two example lasso problems, the left plots show the rejection areas of the dome test and ST3. The upper and lower curved boundaries correspond to Q_u and Q_l, respectively. The black dots are the projections of all codewords $\mathbf{b}_i$ in the projected space $(\mathbf{b}_*^T\mathbf{b}_i, \mathbf{x}^T\mathbf{b}_i)$. The right plots show the total computation time (screening+lasso) of each test for three lasso solvers, on an Intel Xeon X5570 2.93GHz processor. The dome test significantly improves the rejection percentage and the computation speed. Similar behavior is observed on COIL36 and pnoise.

first two data sets.

To get more insight into the tests, we selected a representative lasso problem with $\lambda_{\max} = 0.8$ and $\lambda = 0.5$ in each of the YALEBXF and MNIST500 data sets and for these problems plot in Figure 2 the rejection areas of the dome test and ST3 in the projected space $(\mathbf{b}_*^T\mathbf{b}_i, \mathbf{x}^T\mathbf{b}_i)$, along with the actual projections of the codewords. We see that although the dome test only enlarges the rejection band of ST3 moderately, it can reject an extra 6%-10% of the total codewords. We also compared the total time needed to execute each screening test *and* solve the reduced lasso problem. We report results using three state-of-the-art lasso solvers: Gauss-Seidel, Grafting and Feature-Sign [7, 8]. The dome test results in significant computational savings regardless of the solver used. For the first example, dome test offers 76X-86X speed-up over the original lasso problem without screening and 3.7X-3.9X speed-up over using the test ST3 in [5]. The test itself is very fast: the actual time spend on screening is less than 0.3% of the total computation time in all cases.

5. CONCLUSION

We have investigated an efficient screening test, based on the knowledge that the optimal dual solution $\tilde{\boldsymbol{\theta}}$ is within a region $\mathcal{R}$, that correctly predicts a codeword $\mathbf{b}_i$ to have coefficient $\tilde{w}_i = 0$. The tightness of the bound $\mathcal{R}$ determines the tradeoff between the computational efficiency and the rejection power of the test. In particular, the dome test bounds $\tilde{\boldsymbol{\theta}}$ within a region $\mathcal{G}$ determined by two simple constraints. This test only uses the $2m$ correlations $\{\mathbf{x}^T\mathbf{b}_i, \mathbf{b}_*^T\mathbf{b}_i\}_{i=1}^m$ where $\mathbf{b}_*$ is a vector in $\{\pm\mathbf{b}_i\}_{i=1}^m$ maximizing $\mathbf{x}^T\mathbf{b}_*$. The test can be executed in time linear in the number of codewords. All previous tests in the literature are $\mathcal{R}$ region tests with $\mathcal{G} \subset \mathcal{R}$. Therefore our test is stronger than previous tests. Our experimental results confirm the power of the new test. When applied to natural data sets with large $\lambda_{\max}$, it increases rejection percentages by up to 10%, and reduces computation time by a factor of up to 3.9X over the best existing test. For data sets with a wide spread of $\lambda_{\max}$, the test continues to outperform existing tests but differences in rejection percentages and speed-up are less dramatic. This is expected since the test relies on finding a codeword in the dictionary that is highly correlated/anti-correlated with the test vector.

6. REFERENCES

[1] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2008.

[2] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Advances in Neural Information Processing Systems*, 2009, vol. 3.

[3] L.E. Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination in sparse supervised learning," *Arxiv preprint arXiv:1009.3515*, 2010.

[4] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R.J. Tibshirani, "Strong rules for discarding predictors in lasso-type problems," *Arxiv preprint arXiv:1011.2234*, 2010.

[5] Z. J. Xiang, H. Xu, and P. J. Ramadge, "Learning sparse representations of high dimensional data on large scale dictionaries," in *Advances in Neural Information Processing Systems*, 2011.

[6] Z. J. Xiang and P. J. Ramadge, "Screening tests for lasso problems," Tech. Rep., Princeton University, 2012.

[7] Mark Schmidt, Glenn Fung, and Rómer Rosales, "Fast optimization methods for L1 regularization: a comparative study and two new approaches," *Machine Learning: ECML 2007*, vol. 4701, pp. 286–297, 2007.

[8] H. Lee, A. Battle, R. Raina, and A.Y. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems*, 2007, vol. 19, p. 801.