Douglas-Rachford Splitting for Cardinality Constrained Quadratic Programming

Enzo Busseti, Hamid Javadi, Reza Takapoui

1 Introduction

In this report, we study the class of Cardinality Constrained Quadratic Programs (CCQP), problems with (not necessarily convex) quadratic objective and cardinality constraints. Many practical problems of importance can be formulated as CCQPs. Examples include sparse principal component analysis [1], [2], cardinality constrained mean-variance portfolio selection problem [3]–[5], subset selection problem in regression analysis [6], and many more.

Since these problems are NP-hard in general, branch-and-bound type methods are usually used to solve them. In [7], a tailored branch-and-bound implementation with pivoting algorithm is introduced. In [8], a local relaxation algorithm is discussed which is based on sequence of small, local, quadratic-programs.

In this report we focus on approximately solving CCQPs using Douglas-Rachford (DR) splitting. Douglas-Rachford splitting is an operator splitting method which was introduced in 1970s, but attracted a lot of attention after its close relation with alternating direction method of multipliers was shown [9]. Even though it is originally designed to solve convex optimization problems, recently, using ideas from semi analytic geometry [10], some convergence results for the noncovex case was shown [11]–[13]. Also this method shows competitive performance in practice.

The rest of this report is as follows: in §2 we introduce a general framework for CCQPs that we study. In §3 and §4 we discuss two important examples, finding a sparse solution to a set of linear equations, and sparse principal component analysis (SPCA). In §5 discuss future work and §6, Appendix, includes some of the proofs that we omitted from this report.

2 Problem statement

2.1 Problem

We study the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C}, \end{array} \tag{2.1}$$

where f is a quadratic function and $\mathcal{C}$ is a semi-algebraic set. We assume that the projection onto $\mathcal{C}$ can computationally efficiently be done via a projection function Π.

Many problems of interest can be formulated as (2.1). Examples include sparse PCA and finding sparse solutions to underdetermined systems of linear equations. Since many of these problems are NP-hard in general, no polynomial algorithm is guaranteed to solve these problems. However, many heuristics are proposed to approximately solve this problems (*i.e.*, finding an approximate solution.)

2.2 Algorithm

We will discuss the performance of Douglas-Rachford (DR) algorithm on problem (2.1). Each iteration of this algorithm is depicted in Algorithm 1. We assume that $f : \mathbb{R}^n \to \mathbb{R}^n$ is given by $f(x) = (1/2)x^T Px + q^T x$. The first step involves the proximal step for a quadratic function and is in fact a linear operator. In order for this step to be well-defined, we require that $1 + \gamma\lambda_{\text{min}}(P) > 0$. If the step size γ doesn't change in the algorithm, it is computationally beneficial to cache the factorization of the matrix in this step. The operator Π in the second step is projection into $\mathcal{C}$. The third step is simply a dual update. In the following sections, we explain when and why we need to update

Algorithm 1 DR splitting for problem (2.1)

```
Input: $P \in \mathbb{R}^{n \times n}$, $q \in \mathbb{R}^n$, and $k \in \mathbb{Z}_+$
Initialize: Initial point $x^0$, parameter $\gamma > 0$, iterate $t = 0$, and $f^{\text{best}} = \infty$
 1: $y^{t+1} := -\left(\gamma P + I\right)^{-1}\left(\gamma q + x^t\right)$.
 2: $z^{t+1} := \Pi\left(2y^{t+1} - x^t\right)$.
 3: $x^{t+1} := x^t + z^{t+1} - y^{t+1}$.
 4: update $\gamma$ if needed.
 5: $t := t + 1$
 6: if $f(z^t) < f(z^{\text{best}})$ then
 7:     $z^{\text{best}} = z^t$, $f^{\text{best}} = f(z^t)$.
 8: if Convergence criterion is not met then
 9:     Go to step 1.
10: output $z^{\text{best}}$.
```

the parameter γ. Also, since this algorithm is not necessarily a descent algorithm, it is critical to keep track of the best point found by the algorithm.

2.3 Convergence results

The following theorem is a consequence of theorems $1-4$ from [13].

Theorem 1. *Let λ_{min} denote the smallest eigenvalue of P, and $(\cdot)_+ = \max\{\cdot, 0\}$. If throughout the algorithm*

$$(1+\gamma\|P\|_2)^2 - \frac{5\gamma(\lambda_{min})_+}{2} - \frac{3}{2} < 0,$$

and also the sequence $\{(y^t, z^t, x^t)\}$ generated by algorithm 1 is bounded, then $\{(y^t, z^t, x^t)\}$ converges to $\{(y^*, z^*, x^*)\}$, where z^* is a feasible and stationary point of (2.1).

3 Sparse solution to a set of linear equations

3.1 Problem statement

In this section, we are interested in finding a sparse solution to a set of linear equations. Consider the following feasibility problem

$$\begin{array}{ll} \text{find} & x \\ \text{subject to} & \mathbf{card}\, x \le k \\ & Ax = b, \end{array} \qquad (3.1)$$

with decision variable $x \in \mathbb{R}^n$. The problem data here are $k \in \mathbb{Z}_+$, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. We assume that there exists at least one feasible point to this problem.

This is a well known problem which has numerous applications in sparse design, feature selection, and compressed sensing [14], [15].

3.2 Algorithm

We can rewrite (3.1) in equivalent form

$$\begin{array}{ll} \text{minimize} & d(x, \mathcal{A})^2 \\ \text{subject to} & \mathbf{card}\, x \le k, \end{array} \qquad (3.2)$$

where $\mathcal{A} = \{u | Au = b\}$ is the set of solutions to the linear equations, and d is the Euclidian distance. (By equivalence, we mean that two problems have the same set of minimizers, since 3.2 has at least one solution.) Here, the quadratic objective is $f(x) = \|A^T(AA^T)^{-1}Ax - A^T(AA^T)^{-1}b\|^2$. Also, the second step is the projection into the set of points with cardinality less than or equal to k. This step consists of finding the k largest elements of the point in absolute value and replacing other elements with zero.

3.3 Convergence

The following lemma asserts that under some conditions on the matrix A the sequence $\{(y^t, z^t, x^t)\}$ generated by algorithm 1 is bounded.

Lemma 3.1. *If throughout the algorithm, $\gamma \ge \sqrt{3/2} - 1$ is bounded and for any v in null space of A, and any $\Lambda \subseteq \{1, \dots, n\}$ with $|\Lambda| \le k$,*

$$\sum_{i\in\Lambda} h_i^2 < 0.5(1-c)\|h\|_2^2,$$

for some $0 < c \le 1$, then the sequence $\{(y^t, z^t, x^t)\}$ generated by the algorithm 1 is bounded.

Proof. The proof can be found in Appendix A.1. □

From §6 of [16], we know that if $m \in \Omega(k\log(n/k))$, then random Gaussian matrices $A \in \mathbb{R}^{m\times n}$ satisfy this condition, with high probability.

Combining lemma 3.1 with theorem 1 we can prove the following result.

Theorem 2. *Under the assumptions of lemma 3.1, and the additional assumption that $\gamma = \sqrt{3/2} - 1$, then the sequence $\{z^t\}$ generated by algorithm 1 converges to a stationarty point x^*.*

Remark 3.1. Even though theorem 2 proves the convergence for $\gamma = \sqrt{3/2} - 1$, in practice, we observe that in case of Gaussian matrices A, the algorithm converges for larger parameters γ. In fact, in the next part we will see that the fixed points of the algorithm for $\gamma > 1$ have useful properties.

3.4 Quality of stationary points

Theorem 3. *Assume that (y^*, z^*, x^*) is the limit point of algorithm 1 for $\gamma > 1$. Let $\tilde{x}$ be the projection of x^* onto $\{u | Au = b\}$. Then $\tilde{x} - \Pi\tilde{x}$ is in range of A^T and moreover*

$$\|\tilde{x} - \Pi\tilde{x}\|_\infty \le \frac{1}{\gamma} \min_{i\in\mathrm{supp}(\Pi\tilde{x})} |(\Pi\tilde{x})_i|.$$

Conversely, from any point $\tilde{x}$ satisfying these properties, we can construct fixed point (y^*, z^*, x^*) to the algorithm by

$$\begin{array}{rcl} y^* & = & \dfrac{\gamma}{1+\gamma}\Pi\tilde{x}, \\ z^* & = & \dfrac{\gamma}{1+\gamma}\Pi\tilde{x}, \\ x^* & = & \Pi\tilde{x} + \gamma(\Pi\tilde{x} - \tilde{x}). \end{array}$$

Proof. The proof can be found in Appendix A.2. □

3.5 Numerical results

We generate a series of problems for various values of m, n, and k. In each problem, we generate a random Gaussian matrix $A \in \mathbb{R}^{m\times n}$ and choose b such that the problem is feasible. We run the algorithm 1 and report the normalized error $\frac{\|x-x^\star\|_2}{\|x^\star\|_2}$ (similar to [17]). Each experiment is repeated for 100 times with $\gamma = 10$. We plot the normalized error in figure 3.5.

We compare the results to the well-know ℓ_1 heuristic. In this heuristic, the following relaxed convex problem is solved

$$\begin{array}{ll} \text{minimize} & \|x\|_1 \\ \text{subject to} & Ax = b. \end{array} \tag{3.3}$$

We plot the normalized error for this heuristic in Figure 2. We see that smaller normalized error is achieved by using Douglas Rachford splitting with $\gamma = 10$.

4 Sparse PCA

4.1 Problem statement

Sparse principal component analysis is a well known problem which has numerous applications in different areas [1], [2]. This problem is formulated as the following

$$\begin{array}{ll} \text{maximize} & \frac{1}{2}x^T P x \\ \text{subject to} & \mathbf{card}\, x \le k \\ & \|x\|_2 = 1. \end{array} \tag{4.1}$$

Without loss of generality we can assume that $P \in \mathbb{R}^{n\times n}$ is symmetric and $\lambda_{\max}(P) = 1$. Different algorithms have been proposed in the literature for (approximately) solving this problem. One of these algorithms which works well in practice is an iterative method known as *truncated power method* [18]. This method iteratively applies matrix P and projects onto the feasible set of problem (4.1) to find an approximate solution. Although this method works well in practice there is no guarantee for its convergence for a general matrix P.

4.2 Algorithm

We use the Douglas-Rachford Algorithm 1 to solve the problem (4.1). We notice that the second step of the algorithm (projection onto $\mathcal{C}$) is computationally efficient.

Proposition 4.1. *The projection of $x \in \mathbb{R}^n$ on the set $\mathcal{C} = \{x \in \mathbb{R}^n; \|x\|_2 = 1, \mathbf{card}\, x \le k\}$ is given by*

$$\Pi(x) = \frac{\pi_1(x)}{\|\pi_1(x)\|}$$

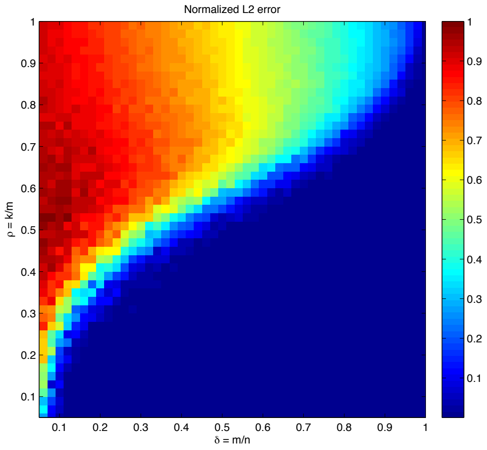

Fig. 1. Average (over 100 independent experiments) of the normalized ℓ_2 error for the Douglas-Rachford splitting method. On the x axis we vary m/n from 0 to 1. On the y axis we vary k/m from 0 (bottom) to 1 (up).

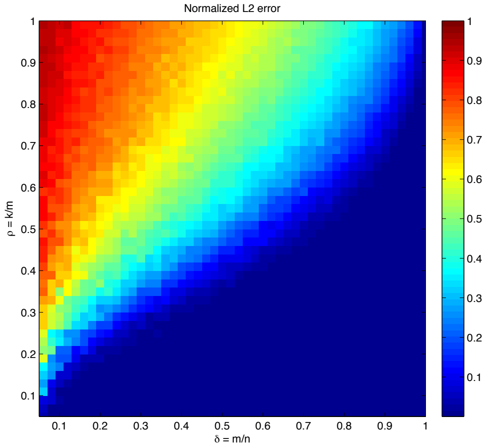

Fig. 2. Average (over 100 independent experiments) of normalized ℓ_2 error for the ℓ_1 relaxation method. On the x axis we vary m/n from 0 to 1. On the y axis we vary k/m from 0 (bottom) to 1 (up).

where $\pi_1(x)$ is the projection of x onto $\{x \in \mathbb{R}^n; \mathbf{card}\, x \le k\}$ and is found by keeping the k largest entries of x in absolute value and zeroing out other entries.

Proof. The proof can be found in Appendix A.3. □

4.3 Convergence

The following lemma shows that the sequence generated by algorithm 1 for sparse PCA problem is bounded under milder assumptions.

Lemma 4.2. *Suppose that in algorithm 1, $\gamma < 0.5$. Then the sequence $\{(y^t, z^t, x^t)\}$ generated by this algorithm is bounded.*

Proof. The proof can be found in the Appendix A.4. □

Now we can use theorems $1 - 4$ in [13] and lemma 4.2 to derive the following theorem.

Theorem 4. *Suppose that the parameter $\gamma > 0$ is chosen such that*

$$(1 + \gamma L)^2 + \frac{5}{2}\gamma - \frac{3}{2} < 0,$$

where $L = \max\{1, -\lambda_{min}(P)\}$. Then the sequence sequence $\{z^t\}$ generated by algorithm 1 converges to a stationary point for problem 4.1.

4.4 Numerical results

We compare the Douglas-Rachford splitting method on this problem with some other competing algorithms: the truncated power method of [18], the ADMM of [9] and a "naive" solution (sparse approximation of the eigenvector of P with largest eigenvalue). The ADMM iterates are described in Appendix B. In our simulations the matrix P is generated according to

$$P = \beta z_0 z_0^T + W,$$

where $\mathbf{card}\, z_0 \leq k$, $\|z_0\|_2 = 1$ and W is a random Wigner matrix normalized such that $\|W\| = 1$. In addition, β is a parameter denoting signal to noise ratio of the model. The results can be seen in Figures 3 and 4. Unfortunately, we note that the Douglas-Rachford splitting method does not achieve noticeably better performance than the Truncated Power method of [18]. Our implementation of the ADMM algorithm achieves also similar results. However, compared to those two methods, we proved that the Douglas-Rachford method has somewhat better convergence properties (Theorem 4). In other words, DR splitting method converges under the conditions of theorem 4, whereas the truncated power method does not converge in the most of examples so we have to keep track of the best point. Another important fact that we observed in the simulations was that the quality of the achieved solution depends highly on the starting point in both methods (Truncated power and DR). For the results that are depicted in Figures 3, 4 we have used the leading eigenvector of P as the starting point for both methods.

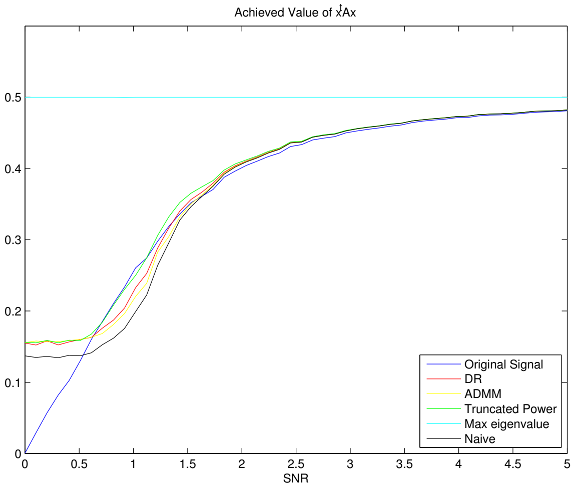

Fig. 3. Average (over 100 independent experiments) of $(z^*)^T A z^*$ for the optimal solution z^*, as a function of the signal to noise ratio β.

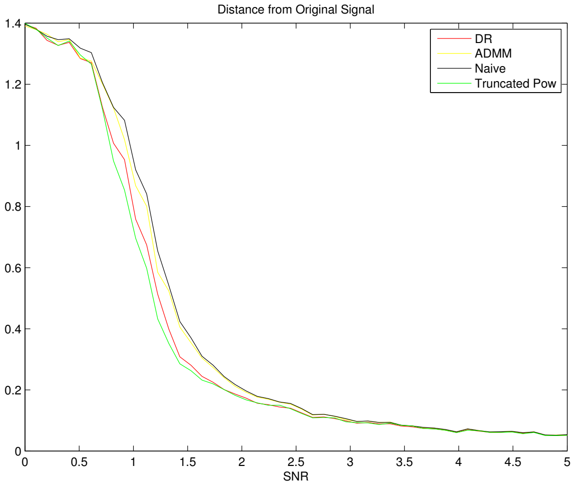

Fig. 4. Average (over 100 independent experiments) distance of the optimal solution from the true solution $\|z^* - z_0\|_2$, as a function of the signal to noise ratio β.

5 Future work

We have observed that in practice, for Gaussian matrices A, the result of Theorem 2 is true even for larger values of γ. We will try to investigate this more. We are also interested in theoretically quantifying the suboptimality of fixed points of DR for sparse PCA based on the starting point. We also plan to analyze a wider range of examples from CCQPs with DR splitting.

References

[1] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[2] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet, "A direct formulation for sparse pca using semidefinite programming," *SIAM review*, vol. 49, no. 3, pp. 434–448, 2007.

[3] T.-J. Chang, N. Meade, J. E. Beasley, and Y. M. Sharaiha, "Heuristics for cardinality constrained portfolio optimisation," *Computers & Operations Research*, vol. 27, no. 13, pp. 1271–1302, 2000.

[4] D. Li, X. Sun, and J. Wang, "Optimal lot solution to cardinality constrained mean–variance formulation for portfolio selection," *Mathematical Finance*, vol. 16, no. 1, pp. 83–101, 2006.

[5] D. X. Shaw, S. Liu, and L. Kopman, "Lagrangian relaxation procedure for cardinality-constrained portfolio optimization," *Optimisation Methods & Software*, vol. 23, no. 3, pp. 411–420, 2008.

[6] A. Miller, *Subset selection in regression*. CRC Press, 2002.

[7] D. Bertsimas and R. Shioda, "Algorithm for cardinality-constrained quadratic optimization," *Computational Optimization and Applications*, vol. 43, no. 1, pp. 1–22, 2009.

[8] W. Murray and H. Shek, "A local relaxation method for the cardinality constrained portfolio optimization problem," *Computational Optimization and Applications*, vol. 53, no. 3, pp. 681–709, 2012.

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, pp. 1–122, 2010.

[10] J. Bolte, A. Daniilidis, and A. Lewis, "The lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 1205–1223, 2007.

[11] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, 2013.

[12] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the kurdyka-lojasiewicz inequality," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 438–457, 2010.

[13] G. Li and P. T. K, "Douglas-rachford splitting for nonconvex feasibility problems," *arXiv preprint arXiv:1409.8444v2*, 2015.

[14] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *Information Theory, IEEE Transactions on*, vol. 52, no. 2, pp. 489–509, 2006.

[15] D. L. Donoho and J. Tanner, "Sparse nonnegative solution of underdetermined linear equations by linear programming," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 27, pp. 9446–9451, 2005.

[16] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," *Journal of the American mathematical society*, vol. 22, no. 1, pp. 211–231, 2009.

[17] D. Donoho and J. Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1906, pp. 4273–4293, 2009.

[18] X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 899–925, 2013.

Appendix A
Proofs

A.1 Proof of lemma 3.1

Let

$$\begin{aligned} A' &= 2A^{\mathsf{T}}(AA^{\mathsf{T}})^{-1}A, \\ b' &= 2A^{\mathsf{T}}(AA^{\mathsf{T}})^{-1}b. \end{aligned}$$

Using algorithm 1, we can see that the sequence x^t is generated according to

$$x^{t+1} = \frac{1}{2}\left(\left(I + D^t - \frac{\gamma D^t A'}{1+\gamma}\right)x^t + \frac{\gamma D^t b'}{1+\gamma}\right),$$

where D^t is a diagonal matrix with k, $+1$'s and $n-k$, -1's on the diagonal. Therefore, it is sufficient to show that the maximum singular value of matrix $B^t = \frac{1}{2}\left(I + D^t - \frac{\gamma D^t A'}{1+\gamma}\right)$ is less than 1 for all possible D^t's. Let $x = x_\perp + x_\parallel$, $\|x\|_2 = 1$ be an arbitrary vector in $\mathbb{R}^n$; where, $x_\perp$ is in the null space of A and $x_\parallel$ is in the range of A^{T}. We have

$$\begin{aligned} \|B^t x\|_2^2 =& \frac{1}{4}\|x_\parallel\|_2^2 + \frac{(1-\gamma)^2}{4(1+\gamma)^2}\|x_\parallel\|_2^2 + \frac{1}{4}\|x_\perp\|_2^2 \\ &+ \frac{1}{4}\|x_\perp\|_2^2 + \frac{1-\gamma}{2(1+\gamma)}\left\langle x_\parallel, D^i x_\parallel\right\rangle \\ &+ \frac{1}{2}\left\langle x_\perp, D^i x_\perp\right\rangle + \frac{1}{2}\left\langle x_\parallel, D^i x_\perp\right\rangle \\ &+ \frac{1-\gamma}{2(1+\gamma)}\left\langle x_\parallel, D^i x_\perp\right\rangle. \end{aligned}$$

Letting $\|x_\parallel\| = u$, since $\left\langle x_\perp, D^i x_\perp\right\rangle < (2c-1)(1-u^2)$,

$$\begin{aligned} \|B^t x\|_2^2 <& \frac{u^2}{4} + \frac{(1-\gamma)^2 u^2}{4(1+\gamma)^2} \\ &+ \left(1-u^2\right)\left(c - \frac{1}{2}\right) + \frac{|1-\gamma| u^2}{2(1+\gamma)} \\ &+ \frac{1}{1+\gamma}\left\langle x_\parallel, D^i x_\perp\right\rangle. \end{aligned}$$

Using,

$$\left\langle x_\parallel, D^i x_\perp\right\rangle \le u\sqrt{1-u^2},$$

it can be seen that $\|B^t x\|_2 < 1$, for all x such that $\|x\|_2 \le 1$. Therefore the maximum singular value of B^t is less than one and the proof is complete.

A.2 Proof of theorem 3

First let's assume that $\tilde{x}$ has the conditions stated in lemma 3. Therefore we can write

$$\tilde{x} = x^* + \left(1 + \frac{1}{\gamma}\right)\left(\Pi(x^*) - x^*\right), \qquad \text{(A.1)}$$

for some $x^* \in \mathbb{R}^n$. Note that since $\tilde{x} - \Pi(\tilde{x}) \in \mathcal{R}(A^{\mathsf{T}})$, $\Pi(x^*) - x^*$ is also in the range of A^{T}. Thus, multiplying both sides of (A.1) by $A^{\mathsf{T}}\left(AA^{\mathsf{T}}\right)^{-1}A$ gives us

$$\begin{aligned}A^{\mathsf{T}}\left(AA^{\mathsf{T}}\right)^{-1}b =& A^{\mathsf{T}}\left(AA^{\mathsf{T}}\right)^{-1}Ax^* \\ &+\left(1+\frac{1}{\gamma}\right)\left(\Pi(x^*)-x^*\right).\end{aligned}$$

Thus,

$$\begin{aligned}x^*+A^{\mathsf{T}}\left(AA^{\mathsf{T}}\right)^{-1}(b-Ax^*) =& \frac{1}{\gamma}\left(\eta(x^*;k)-x^*\right) \\ &+\eta(x^*;k).\end{aligned}$$

Which means that

$$x^*+\frac{\gamma}{(1+\gamma)}A^{\mathsf{T}}\left(AA^{\mathsf{T}}\right)^{-1}(b-Ax^*)=\Pi(x^*).$$

Therefore, x^* is a fixed point of algorithm 1. Conversely, assume that x^* is a fixed point of algorithm 1. Take

$$\tilde{x}=x^*+A^{\mathsf{T}}\left(AA^{\mathsf{T}}\right)^{-1}(b-Ax^*).$$

It can be easily seen that $A\tilde{x}=b$. In addition, since x^* is a fixed point of algorithm 1, note that we have

$$y^*=z^*=\Pi(2y^*-x^*)=\Pi(x^*)$$

and condition 2 also holds. Finally,

$$\tilde{x}-\Pi(\tilde{x})=cA^{\mathsf{T}}\left(AA^{\mathsf{T}}\right)^{-1}(b-Ax^*),$$

where c is a constant depending on γ. Therefore condition 3 also holds and the proof is complete.

A.3 Proof of Proposition 4.1

By definition, for any $a\in\mathbb{R}^n$; $\Pi(a)$, is the optimal solution to this problem

$$\begin{array}{ll}\text{minimize} & \|x-a\|_2^2 \\ \text{subject to} & \|x\|_0\le k \\ & \|x\|_2=1.\end{array}\qquad\text{(A.2)}$$

Note that for every x which is feasible for (A.2)

$$\begin{aligned}\|x-a\|_2^2&=\|x\|_2^2-2x^Ta+\|a\|_2^2\\&=-2x^Ta+1+\|a\|_2^2.\end{aligned}$$

Since $\|a\|_2^2$ is a constant, $\|x\|_2=1$, optimal solution of (A.2) is the same as the optimal solution of the following problem

$$\begin{array}{ll}\text{maximize} & \cos\theta_{x,a} \\ \text{subject to} & \|x\|_0\le k \\ & \|x\|_2=1.\end{array}\qquad\text{(A.3)}$$

Obviously, a in (A.3) can be normalized and the problem will remain the same. Thus, (A.3) is equivalent to

$$\begin{array}{ll}\text{maximize} & \cos\theta_{x,\hat{a}} \\ \text{subject to} & \|x\|_0\le k \\ & \|x\|_2=1.\end{array}\qquad\text{(A.4)}$$

In (A.4), $\hat{a}=a/\|\pi_1(a)\|_2$; where $\pi_1(a)$ is the projection of a onto $\mathcal{C}'=\{u|\,\mathbf{card}\,u\le k\}$. Now, by the same way as we established that the problems (A.2), (A.3) have the same optimal solutions, we can argue that the optimal solution of this problem is the same as of the following problem

$$\begin{array}{ll}\text{minimize} & \|x-\hat{a}\|_2^2 \\ \text{subject to} & \|x\|_0\le k \\ & \|x\|_2=1.\end{array}\qquad\text{(A.5)}$$

Now, by relaxing this problem we can say that the optimal value of (A.5) is bigger than the optimal value of the following problem

$$\begin{array}{ll}\text{minimize} & \|x-\hat{a}\|_2^2 \\ \text{subject to} & \|x\|_0\le k.\end{array}\qquad\text{(A.6)}$$

It can be easily seen that the optimal solution of (A.6) is

$$x^*=\pi_1(\hat{a})=\frac{\pi_1(a)}{\|\pi_1(a)\|_2}.$$

But x^* is feasible for (A.5). Therefore, x^* is the optimal solution of (A.5). Therefore

$$\Pi(a)=\frac{\pi_1(a)}{\|\pi_1(a)\|_2}.$$

Therefore, projection onto the feasible set of problem (4.1) can be done easily and the Douglas-Rachford splitting algorithm is computationally feasible.

A.4 Proof of proposition 4.2

It can be easily seen that z^t is bounded. Therefore it suffices to show that for all t, x^t is bounded. To prove this, note that

$$x^{t+1}=Bx^t+v^t,$$

where $B=I-(I-\gamma A)^{-1}$, $\|v^t\|_2=1$. Therefore

$$x^{t+1}=B^{t+1}x^0+\sum_{i=0}^{t}B^{t-i}v^i.$$

Therefore, using $\|x^0\|_2=1$, $\|v^i\|_2=1$ we have

$$\|x^{t+1}\|_2\le\sum_{i=0}^{\infty}|\lambda_{\max}(B)|^i,\qquad\text{(A.7)}$$

where $\lambda_{\max}(B)$ is the largest eigenvalue of B in absolute value. Note that

$$\frac{-\gamma}{1-\gamma} \leq \lambda_i(B) \leq \frac{\beta\gamma}{1+\beta\gamma}.$$

Therefore if $\gamma < 1/2$, $|\lambda_{\max}(B)| < 1$, and using (A.7), $\|x^{t+1}\|_2$ is bounded. Using this, it can be easily seen that $\|y^{t+1}\|_2$ will be bounded and the proof is complete.

Appendix B
ADMM for Sparse PCA

Consider the problem of equation 4.1. We rewrite it in the form

$$\begin{array}{rl} \text{minimize} & f(x) + h(z) \\ \text{s.t.} & x = z \end{array}$$

where

$$f(x) = -\frac{1}{2}x^T A x$$

$$h(z) = I_{\|x\|_2=1} + I_{\|x\|_0 \leq k}.$$

We define the augmented Lagrangian

$$L(x, z, \nu) = f(x)+h(z)+\nu^T(x-z)+(\rho/2)\|x-z\|_2^2,$$

we apply the ADMM algorithm [9] to this non-convex problem. The updates are:

$$x^+ = \arg\inf_x L(x, z, \nu)$$

$$z^+ = \arg\inf_z L(x^+, z, \nu)$$

$$\nu^+ = \nu + \rho(x^+ - z^+).$$

The x update becomes

$$x^+ = \arg\inf_x(\frac{1}{2}x^T(\rho I - A)x + (\nu - \rho z)^T x)$$

which has solution only if $\rho \geq \lambda_{max} A$, and we get

$$x^+ = (\rho I - A)^{-1}(\nu - \rho z).$$

For the z update instead we get

$$z^+ = \arg \sup_{\|z\|_2=1, \|z\|_0 \leq k} (\nu+\rho x^+)^T z = \Pi(\nu+\rho x^+)$$

which by Proposition 4.1 is equal to

$$z^+ = \frac{\pi_1(\nu + \rho x^+)}{\|\pi_1(\nu + \rho x^+)\|_2},$$

where $\pi_1(\cdot)$ is the projection onto $\{x \in \mathbb{R}^n | \,\mathbf{card}\, x \leq k\}$.