Solving the L_1 regularized least square problem via a box-constrained smooth minimization

Majid Mohammadi, Wout Hofman, Yaohua Tan and S. Hamid Mousavi

***Abstract*—In this paper, an equivalent smooth minimization for the L_1 regularized least square problem is proposed. The proposed problem is a convex box-constrained smooth minimization which allows applying fast optimization methods to find its solution. Further, it is investigated that the property "the dual of dual is primal" holds for the L_1 regularized least square problem. A solver for the smooth problem is proposed, and its affinity to the proximal gradient is shown. Finally, the experiments on L_1 and total variation regularized problems are performed, and the corresponding results are reported.**

***Index Terms*—sparse, L_1 regularization, smooth, total variation, proximal gradient**

I. Introduction

Finding the solution to the least square problem with L_1 regularization is of utmost importance due to its myriad applications including but not limited to pattern recognition [30], [31], feature selection [33], [35], image processing [8], [34] and bioinformatics [12], [26]. The minimization to obtain such a solution is

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 \tag{1}$$

where $x \in R^l$ is the coefficient vector, $A \in R^{n\times l}$ is a basis, $b \in R^{l\times 1}$ is a regression vector, λ is the non-negative regularization parameter, and $\|.\|_p$ is the p-norm.

The challenge of finding the optimal solution of minimization (1) is its non-smoothness, thereby impeding us to leverage fast optimization methods. Therefore, developing efficient algorithms to find the optimal solution of this minimization has drawn a lot of attentions in the recent decade.

One approach to solving the problem is to use two auxiliary variables into which x is split [10]. The resultant problem, though smooth, has the double dimension with respect to the minimization (1) so that obtaining its solution becomes more time- and memory-consuming. Another approach is to utilize the dual of the minimization (1), which is a smooth box-constrained smooth problem [15], [36]. One main drawback is the calculation of the primal solution x, which needs the computation of $(A^TA)^{-1}$. In general sense, however, A^TA is not necessarily nonsingular; therefore it is not invertible. Even if it is invertible, it would be prolonged to compute $(A^TA)^{-1}$ for large-scale problems. Another strategy to solve the problem (1) is based on the subgradient. However, subgradient-based methods are too slow in comparison with the gradient based techniques which makes it inappropriate in the case of large-scale problems.

In this letter, we firstly utilize the dual problem of the minimization (1) and show that the property "the dual of dual is primal" is correct for the minimization (1). Such a property holds for all convex linear programming but does not generally hold for all convex nonlinear problems. We further derive a smooth problem with box constraints which is equivalent to the minimization (1). Moreover, a solver for the derived smooth problem is proposed, and its affinity with the proximal gradient is demonstrated. We initially suppose that A^TA is invertible, but neither the resultant smooth problem nor the proposed solver would be predicated on the invertibility of A^TA. The smooth minimization is further adjusted for the total variation-regularized problems.

This letter is organized as follows. The smooth equivalent problem is derived in Section II, and its generalization to the total variation is brought in Section III. A solver is proposed in Section IV and its relation with the proximal gradient is investigated. Finally, Section V includes the experiments corresponding to the L_1 and total variation-regularized problems.

M. Mohammadi and Yaohua Tan are with the Faculty of Technology, Policy and Management, Delft University of Technology, Netherlands. E-mail: m.mohammadi@tudelft.nl and y.tan@tudelft.nl.

Wout Hofman is a senior researcher at TNO, Netherlands. E-mail: wout.hofman@tno.nl.

S. H. Mousavi is with the Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Carl von Ossietzky University of Oldenburg, Germany. E-mail: hamid.mousavi@uni-oldenburg.de.

II. Smooth equivalent problem

In this section, we firstly introduce the dual of the minimization (1) and prove that the dual of dual is primal. Then, a smooth equivalent problem to minimization (1) is obtained.

A. The dual of dual is primal

As the minimization (1) is not smooth, its dual problem cannot be derived immediately. However, the dual problem can be obtained by a change of variable and using a theorem in [4] as (see [29] for the whole procedure):

$$\begin{aligned}\min_z\ & \frac{1}{2}z^T(A^TA)^{-1}z + z^T(A^TA)^{-1}A^Tb\\ s.t.\ & \quad -\lambda \le z_i \le \lambda \quad i = 1,2,...,l.\end{aligned} \tag{2}$$

Further, the relation between the primal and dual variables is

$$x = (A^TA)^{-1}(A^Tb - z). \tag{3}$$

The minimization (2) is smooth and convex and can be solved efficiently by convex minimization methods. Afterward, the solution z is replaced in Eq. (3) to obtain the primal solution x.

Now, the following theorem indicates that the dual of the minimization (2) is the minimization (1).

Theorem 1: Let u and v be the Lagrangian multipliers for the minimization (2). Then
(i) the solution of the primal problem (1) is the difference between the Lagrangian multipliers u, v in the dual problem (2), i.e. $x = u - v$,
(ii) The property "the dual of dual is primal" holds for the minimization (1).

Proof: (i) Let u and v be the Lagrangian multipliers for the problem (2). According to the K.K.T conditions, we have

$$(A^TA)^{-1}z - (A^TA)^{-1}A^Tb + u - v = 0. \tag{4}$$

Replacing $z = A^Tb - A^TAx$ from the Eq. (3) in the above equation, we have $x = u - v$.
(ii) To get the dual, the Lagrangian of the dual problem is firstly written as

$$\begin{aligned} J = \frac{1}{2}z^T(A^TA)^{-1}z - z^T(A^TA)^{-1}A^Tb \\ +u^T(z-\lambda) + v^T(-\lambda - z) \end{aligned}$$

Since $z = A^Tb - (A^TA)(u - v)$ based on Eq. (4), the Lagrangian function can be rewritten as

$$\begin{aligned} &J = \frac{1}{2}(u-v)^TA^TA(u-v) + \frac{1}{2}(u-v)^TA^Tb \\ &-\frac{1}{2}b^TA(u-v) + u^T(A^Tb - (A^TA)(u-v) - \lambda) \\ &+ v^T(-\lambda - A^Tb + A^TA(u-v)) \\ &\Rightarrow J = -\frac{1}{2}(u-v)^TA^TA(u-v) + (u-v)^TA^Tb - \lambda(u-v). \end{aligned}$$

Since $x = u - v, u, v \geq 0$ and $uv = 0$ according to K.K.T conditions of the dual problem (2), we have $\|x\|_1 = u + v$. Hence, the dual problem based on the above equation is

$$\max_x -\|Ax - b\|_2^2 - \lambda\|x\|_1 \tag{5}$$

which is equivalent to the minimization (1) and the proof is complete. ■

B. Smooth problem

The smooth equivalent problem to the minimization (1) can be easily obtained by replacing z in the dual problem. Thanks to Eq. (3), we have

$$z = A^Tb - A^TAx. \tag{6}$$

Replacing (6) into the minimization (2) and doing some calculus, we obtain

$$\begin{aligned} &\min_x x^TA^TAx \\ &s.t. \quad -\lambda \leq A^TAx - A^Tb \leq \lambda. \end{aligned} \tag{7}$$

The problem (7) is convex with box constraints. As it is smooth, the optimization methods for constrained problems can be applied to find its solution. In contrast to the dual problem, there is no need to compute the inverse matrix $(A^TA)^{-1}$ for solving this problem. Further, it directly obtains the solution that is desired, e.g. x.
In further sections, a new method is proposed for solving the minimization (7) and its affinity with the proximal gradient is also investigated.

III. Total variation regularization

Another important non-smooth problem is the total variation regularized problem which can be written as L_1 regularized least square as

$$\min_u \|y - u\|_2^2 + \lambda\|Du\|_1 \tag{8}$$

where $y, u \in R^l$ and $D \in R^{l-1,l}$ is defined as

$$D = \begin{bmatrix} 1 & -1 & 0 & \dots 0 & 0 \\ 0 & 1 & -1 & \dots 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -1 \end{bmatrix}.$$

Although this problem is seemingly similar to the minimization (1), finding its solution is more challenging due to the multiplication of a matrix inside the L_1 regularization.
Z. Harchaoui and C. Levy-Leduc [18] transform the minimization (8) to the problem (1). Their results are brought in the following theorem.

Theorem 2 ([18]): The minimizations (1) and (8) are equivalent if the variables in the problem 1 are initialized as

$$\begin{aligned} &x = Du \\ &A = D^T(DD^T)^{-1} \\ &b = D^T(DD^T)^{-1}Dy \end{aligned} \tag{9}$$

where D, y and u are the variables in the problem (8). Furthermore, u is obtained by

$$u = y + D^T(DD^T)^{-1}(x - Dy).$$

Taking into account Theorem 2, the following minimization is equivalent to the problem (8)

$$\begin{aligned} &\min_u \quad u^TD^T(DD^T)^{-1}Du \\ &s.t. \quad -\lambda \leq (DD^T)^{-1}D(u-y) \leq \lambda \end{aligned} \tag{10}$$

IV. A solver and the relation to the proximal gradient method

The minimization (7) is convex, therefore the Karush-Kuhn-Tucker (K.K.T.) conditions are sufficient for optimality [1]. Using the K.K.T conditions, the problem (7) is turned into an equation whose solution is the same as the minimization.

Theorem 3: x is the optimal solution of the minimization (7) if and only if the following equality holds

$$P_\Omega(x - (A^TAx - A^Tb)) = A^Tb - A^TAx \tag{11}$$

where $P_\Omega(.)$ is a piecewise function defined as

$$(P_\Omega(w))_i = \begin{cases} \lambda & w_i > \lambda \\ w_i & |w_i| \leq \lambda \\ -\lambda & w_i < -\lambda \end{cases}$$

Proof: The equation (11) can be easily obtained by writing the K.K.T conditions for the smooth problem (7) (see [2] for details). ■

The equation (11) can be solved by iterative algorithms such as the successive overrelaxation [20], [28]. Another approach is to turn it into a dynamic system whose dynamic equation given by

$$\frac{dx}{dt} = P_\Omega(x - (A^TAx - A^Tb)) - A^Tb + A^TAx. \quad (12)$$

The dynamic system can be seen as a one-layer recurrent neural network and its convergence is guaranteed by a Lyapunov function [14], [19], [32].

It is also possible to solve the minimization (1) through the soft thresholding operator and proximal gradient. The following theorem indicates the solution of the problem (1) using proximal gradient method.

Theorem 4 ([24]): x is the optimal solution of the minimization (1) if and only if it is the fixed point of the following equality

$$P_{prox}(x - (A^TAx - A^Tb)) = x \quad (13)$$

where $P_{prox}(.)$ is the soft thresholding operator defined as

$$(P_{prox}(w))_i = \begin{cases} w_i - \lambda & w_i > \lambda \\ 0 & |w_i| \leq \lambda \\ w_i + \lambda & w_i < -\lambda \end{cases} \quad (14)$$

Although the equalities (11) and (13) are seemingly different in both sides of equations, they bear the same results. The following theorem shows their equality.

Theorem 5: The equation (11) in equivalent to the equation (13).

Proof: We firstly show the equivalence of equations for $x - (A^TAx - A^Tb) > \lambda$. If $x - (A^TAx - A^Tb) > \lambda$, then $P_\Omega(x - (A^TAx - A^Tb)) = \lambda$. Thus, the problem turns into the equation $A^Tb - A^TAx = \lambda$. On top of that, $P_{prox}(x - (A^TAx - A^Tb)) = x - (A^TAx - A^Tb) - \lambda$ and its equation would be $x - (A^TAx - A^Tb) - \lambda = x$. It is readily seen that the proposed solver and the proximal gradient solve the same equation.
Similar equivalences can be obtained for $|x - (A^TAx - A^Tb)| < \lambda$ and $x - (A^TAx - A^Tb) > -\lambda$, and that completes the proof. ■

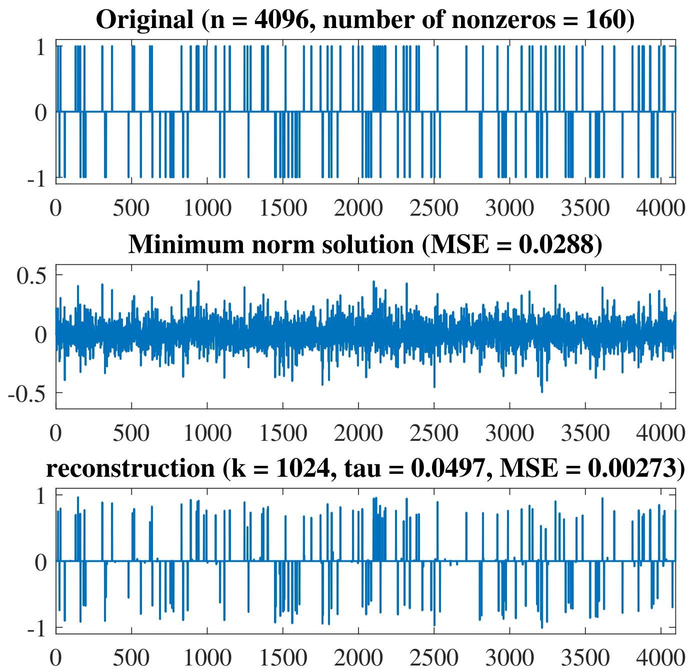
V. Experimental Results

In the section, the experimental results are presented four three different problems. First, a randomly-generated sparse signal is recovered by solving the L_1 regularized least square problem. Then, two applications from total variation based regularization are investigated: the image restoration and the aCGH data recovery.

A. Signal recovery

We generated a sparse signal $x_0 \in R^{4096}$ with 160 spikes; each spike has amplitude $\{-1, 1\}$. This signal is plotted at the top of Fig. 1. A Gaussian noise with the standard variation $\sigma = 0.1$ is added to x_0 to generate the observation y. Then, the measurement matrix $A \in R^{1024\times4096}$ is generated whose entries are i.i.d. according to the standard normal distribution.

Fig. 1: Sparse signal recovery by the proposed method. Top: the randomly generated signal $x_0 \in R^{4096}$ with 160 spikes. Middle: the minimal norm solution obtained by A^Tx_0. Bottom: the recovered signal by the proposed method.



The rows of this matrix are then orthogonalized as done in [5]. The regularization parameter λ is also set as $\lambda = 0.01\|A^Ty\|_\infty$ as any value greater than $\|A^Ty\|_\infty$ for λ leads to the solution zero for the minimization (1) [11], [16]. Given A, y and λ, it is the time estimate x_0 by solving the minimization (1).
The result of the recovery is shown in Fig. 1. The top plot in this figure corresponds to the original noise-free signal which is randomly generated. The bottom plot is the recovered signals by the proposed method, and the middle one is the minimal L_2 norm of the system $Ax = b$. This figure confirms that the recovered signal is faithful in spite of having a few measurements.

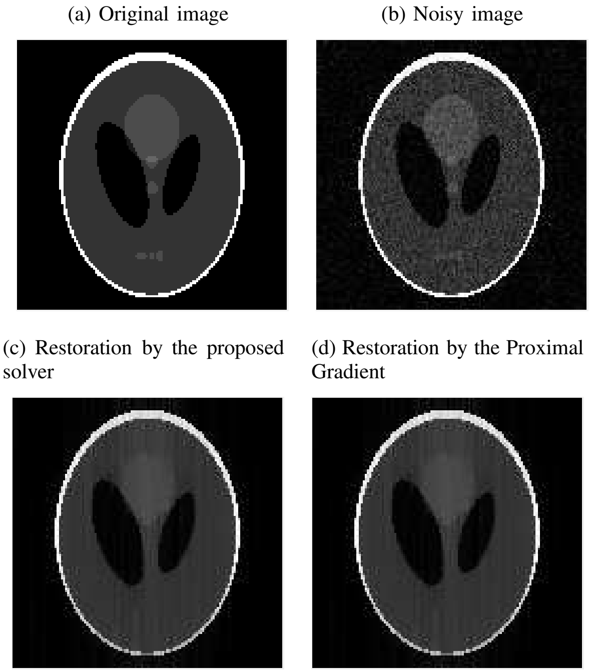
B. Image restoration

The proposed method is applied to the image restoration problem and its result is compared with the proximal gradient. To do so, an MRI image is chosen which is plotted in Fig. 2(a). Further, a random Gaussian noise with $\sigma = 0.05$ is added to the image and the resultant image is in Fig. 2(b). Given the image, the restoration is done by the minimization (8) solved by the proposed method and the proximal gradient. The recovered images are shown in Fig. 2 (c) and (d) for the proposed method and the proximal gradient, respectively.

C. CGH array data recovery

The array comparative genome hybridization (aCGH or CGH array) is a powerful technique to discover the genome-wide DNA copy number variations [25]. However, the experiential aCGH data are highly corrupted by various noises

Fig. 2: The restoration image experiment across an MRI image. (a) the original image, (b) the corrupted image with a noise distributed according to the normal distribution with $\sigma = 0.05$, (c) the recovered image by the proposed method, (d) the recovered image by the proximal gradient.



thereby disabling us to find the change-points from the raw data [9].

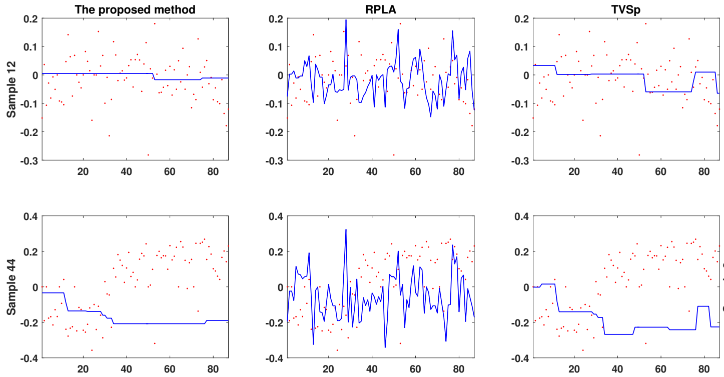
One underlying assumption in aCGH data is that the contiguous chromosome has the identical copy number unless an alteration has happened. Based on this critical assumption, myriad methods have utilized the total variation regularization, whether they process individual samples separately [13], [17], [21] or process multisample data simultaneously [3], [22], [23], [37], [38].

We apply the proposed method on the CGH array from two breast cancer datasets. The Pollack et al. dataset [27] consists of 6691 human mapped genes for 44 primary breast tumors, and the Chin et al. dataset [6] has 2149 clones from 141 primary breast tumors.

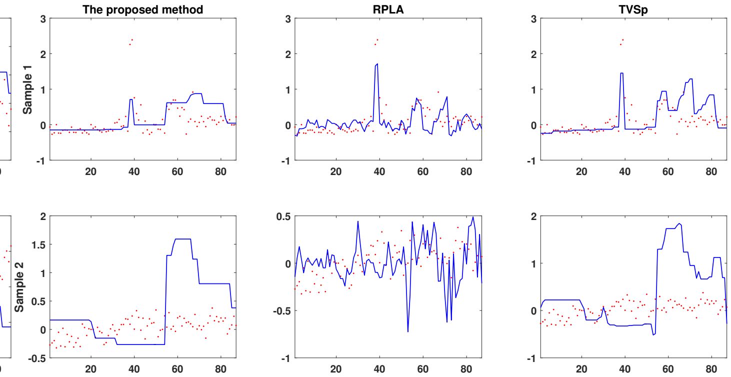
Multiple recovered profiles from the aforementioned datasets are plotted in Figure 3. In this figure, the profiles from the methods TVSp [38] and PLA [37] are also shown. The red dots indicates the raw data and the blue lines are the recovered data by methods. Figure 3 (a) and (b) correspond to two samples from Pollack et al. [27] and Chin et al. [6] datasets, respectively. It is plain to grasp that the proposed method has successfully recovered smooth data from the noisy observations. The recovered profiles by the proposed solver are way smoother than PLA and are competitive with TVSp.

VI. Conclusion

In this article, a smooth problem with box-constraints was shown to be equivalent to the $L1$ regularized least square minimization. A solver proposed for the resultant smooth problem which has affinity with the well-known proximal gradient method. The solver and the smooth problem were further adjusted to encompass the total variation-regularized minimizations, another infamous problem for non-differentiability. For the future work, it is interesting to focus on solving the derived smooth problem more efficiently, and it might be of utmost interest to find the closed-form solution for the $L1$ regularized least square problem.



(a) Selected samples from Pollack et al. dataset [27]



(b) Selected samples from Chin et al. dataset [6]

Fig. 3: Recovered profiles by the proposed method, TVSp [38] and PLA [37]. (a) two samples selected from the Pollack et al. dataset [27]; (b) two samples selected from the Chin et al. dataset [6]. Each row is dedicated to each sample, and each column is devoted to each method. The red dots are the raw observations, and the blue lines are the recovered profiles by each method.

References

[1] Mokhtar S Bazaraa, Hanif D Sherali, and Chitharanjan M Shetty. *Nonlinear programming: theory and algorithms*. John Wiley & Sons, 2013.

[2] Dimitri P Bertsekas and John N Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989.

[3] Kevin Bleakley and Jean-Philippe Vert. The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199*, 2011.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[5] Emmanuel Candes and Justin Romberg. l1-magic: A collection of matlab routines for solving the convex optimization programs central to compressive sampling. *Avaliable: www. acm. caltech. edu/l1magic*, 2006.

[6] Koei Chin, Sandy DeVries, Jane Fridlyand, Paul T Spellman, Ritu Roydasgupta, Wen-Lin Kuo, Anna Lapuk, Richard M Neve, Zuwei Qian, Tom Ryder, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell*, 10(6):529–541, 2006.

[7] Koei Chin, Sandy DeVries, Jane Fridlyand, Paul T Spellman, Ritu Roydasgupta, Wen-Lin Kuo, Anna Lapuk, Richard M Neve, Zuwei Qian, Tom Ryder, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell*, 10(6):529–541, 2006.

[8] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

[9] Lars Feuk, Andrew R Carson, and Stephen W Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, 2006.

[10] Mário AT Figueiredo, Robert D Nowak, and Stephen J Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of selected topics in signal processing*, 1(4):586–597, 2007.

[11] J-J Fuchs. On sparse representations in arbitrary redundant bases. *IEEE transactions on Information theory*, 50(6):1341–1344, 2004.

[12] Xiyi Hang and Fang-Xiang Wu. Sparse representation for classification of tumors using gene expression data. *BioMed Research International*, 2009, 2009.

[13] Jing Hu, Jian-Bo Gao, Yinhe Cao, Erwin Bottinger, and Weijia Zhang. Exploiting noise in array cgh data to improve detection of dna copy number change. *Nucleic acids research*, 35(5):e35, 2007.

[14] Xinjian Huang, Xuyang Lou, and Baotong Cui. A novel neural network for solving convex quadratic programming problems subject to equality and inequality constraints. *Neurocomputing*, 214:23–31, 2016.

[15] Jingu Kim and Haesun Park. Fast active-set-type algorithms for l1-regularized linear regression. In *AISTATS*, pages 397–404, 2010.

[16] S Kim, K Koh, M Lustig, S Boyd, and D Gorinevsky. A method for large-scale l1-regularised least squares problems with applications in signal processing and statistics. *IEEE J. Select. Topics Signal Process*, 2007.

[17] Yoon-ha Lee, Michael Ronemus, Jude Kendall, B Lakshmi, Anthony Leotta, Dan Levy, Diane Esposito, Vladimir Grubor, Kenny Ye, Michael Wigler, et al. Removing system noise from comparative genomic hybridization data by self-self analysis. *arXiv preprint arXiv:1105.0900*, 2011.

[18] Céline Levy-leduc and Zaïd Harchaoui. Catching change-points with lasso. In *Advances in Neural Information Processing Systems*, pages 617–624, 2008.

[19] Qingshan Liu and Jun Wang. A second-order multi-agent network for bound-constrained distributed optimization. *IEEE Transactions on Automatic Control*, 60(12):3310–3315, 2015.

[20] Olvi L Mangasarian and David R Musicant. Successive overrelaxation for support vector machines. *IEEE Transactions on Neural Networks*, 10(5):1032–1037, 1999.

[21] Apratim Mitra, George Liu, and Jiuzhou Song. A genome-wide analysis of array-based comparative genomic hybridization (cgh) data to detect intra-species variations and evolutionary relationships. *PloS one*, 4(11):e7978.

[22] Majid Mohammadi, Ghosheh Abed Hodtani, and Maryam Yassi. A robust correntropy-based method for analyzing multisample acgh data. *Genomics*, 2015.

[23] Hossein Sharifi Noghabi, Majid Mohammadi, and Yao-Hua Tan. Robust group fused lasso for multisample copy number variation detection under uncertainty. *IET Systems Biology*, 10(6):229–236, 2016.

[24] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[25] Daniel Pinkel and Donna G Albertson. Array comparative genomic hybridization and its applications in cancer. *Nature genetics*, 37:S11–S17, 2005.

[26] Roger Pique-Regi, Jordi Monso-Varona, Antonio Ortega, Robert C Seeger, Timothy J Triche, and Shahab Asgharzadeh. Sparse representation and bayesian detection of genome copy number alterations from microarray data. *Bioinformatics*, 24(3):309–318, 2008.

[27] Jonathan R Pollack, Therese Sørlie, Charles M Perou, Christian A Rees, Stefanie S Jeffrey, Per E Lonning, Robert Tibshirani, David Botstein, Anne-Lise Børresen-Dale, and Patrick O Brown. Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, 99(20):12963–12968, 2002.

[28] Zhiquan Qi, Yingjie Tian, and Yong Shi. Successive overrelaxation for laplacian support vector machine. *IEEE transactions on neural networks and learning systems*, 26(4):674–683, 2015.

[29] Mark Schmidt. Two dual problems to l˜ 1 regularized least squares. 2008.

[30] John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang, and Shuicheng Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

[31] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2009.

[32] Youshen Xia, Gang Feng, and Jun Wang. A recurrent neural network with exponential convergence for solving convex quadratic program and related linear piecewise equations. *Neural Networks*, 17(7):1003–1015, 2004.

[33] Allen Y Yang, John Wright, Yi Ma, and S Shankar Sastry. Feature selection in face recognition: A sparse representation perspective. *submitted to IEEE Transactions Pattern Analysis and Machine Intelligence*, 2007.

[34] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

[35] Meng Yang and Lei Zhang. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In *European conference on computer vision*, pages 448–461. Springer, 2010.

[36] Pinghua Gong Changshui Zhang. A fast dual projected newton method for l1-regularized least squares. *Tsinghua University, Beijing*, 2011.

[37] Xiaowei Zhou, Jiming Liu, Xiang Wan, and Weichuan Yu. Piecewise-constant and low-rank approximation for identification of recurrent copy number variations. *Bioinformatics*, page btu131, 2014.

[38] Xiaowei Zhou, Can Yang, Xiang Wan, Hongyu Zhao, and Weichuan Yu. Multisample acgh data analysis via total variation and spectral regularization. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 10(1):230–235, 2013.