ON VECTOR L0 PENALIZED MULTIVARIATE REGRESSION

Akila J. Seneviratne, Student Member, IEEE and Victor Solo, Fellow, IEEE

School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, Australia.

ABSTRACT

The scalar sparse under-determined linear regression problem has had a rapid development with the multivariate version being of more recent interest. In this paper we pose a vector l_0 penalized multivariate regression problem to generate coefficient vectors with shared sparsity profile and then solve the problem with a new cyclic descent algorithm. We give optimality conditions and also discuss penalty parameter selection. Finally we present simulation results that compare our algorithm with alternatives.

Index Terms— multivariate regression, multiple measurement vectors, sparsity, l_o, cyclic descent

1. INTRODUCTION

The task of representing a signal of interest as a linear combination of few elementary signals extracted from a redundant dictionary arises in many applications. Although it is NP hard to find a maximum sparse solution over a general redundant dictionary, many algorithms have shown the ability to recover solutions under certain conditions.

Sparse regression algorithms can be broadly divided in to two types. Algorithms such as forward selection [1] and orthogonal matching pursuit (OMP) [2] are greedy algorithms that iteratively minimize a mean squared error followed by an ad-hoc stopping criterion.

The other type of algorithms solve a penalized or constrained least squares criterion. The l_0 norm penalty promises maximum sparsity however its nonlinear discrete nature poses great difficulty in finding a global minimum. Thus attention has been focussed on convex relaxations of the l_0 norm such as l_1 leading to the LASSO [3], [4]. The FOCUSS algorithm [5] is based on the $l_p(1 > p > 0)$ penalty. Alternatively the l_0 norm is approximated by a differentiable function as in [6].

While multivariate over-determined regression has a long history in statistics [7], [8] the under-determined case has had much less attention particularly in a sparse setting. The sparse version has been motivated by applications such as neuromagnetic inverse problems [9], direction-of-arrival [10], channel equalization [11], and array processing [12] where the multivariate regression problem naturally exists.

A number of sparse scalar regression algorithms have been extended to the multivariate case. Simultaneous orthogonal matching pursuit (SOMP) is presented in [13], [14]. Algorithms that minimize multivariate versions of the penalized or constrained least squares criterion were also developed. [15] and [16] presents algorithms based on the extension of l_1 norm and [17] presents the extension of l_p norm. The vector l_0 norm is approximated by a differentiable zero mean Gaussian function in [10]. The ReMBo algorithm [18] converts the multivariate regression to a scalar regression by randomly combining the measurement vectors. The performance of multivariate regression algorithms have been compared under various conditions [19], [20] but most of the work has been done on noiseless systems.

In this paper we develop a cyclic descent algorithm to minimize a vector l_0 penalized multivariate regression criterion. Cyclic descent is like a classic Gauss-Seidel algorithm as opposed to Landweber based algorithms such as [21] which are like classic Jacobi algorithms.

The remainder of the paper is organized as follows. Section 2 introduces the notation and develops the multivariate l_0 penalized regression problem. Section 3 presents the cyclic descent algorithm followed by a discussion of algorithm initialization, termination, optimality conditions and penalty parameter selection. Section 4 has simulations comparing the performance of the algorithm with that of existing algorithms. Conclusions are in section 5.

2. VECTOR L_0 PENALIZED LEAST SQUARES

Consider the multivariate measurement system :

$$y_{(c)} = X\beta_{(c)} + \varepsilon, \quad c = 1, \ldots, d,$$

where $y_{(c)}$ is a n dimensional measurement vector, $X_{n\times p}$ is a regression matrix or dictionary and $\beta_{(c)}$ is a p dimensional coefficient vector. When $n < p$, X is called a redundant dictionary and the system is under-determined. When d measurement vectors are collected together we can rewrite this as,

$$Y_{n\times d} = X_{n\times p}B_{p\times d} + E, \tag{1}$$

This work was partly supported by an ARC(Australian Research Council) grant.

where $Y_{n\times d} = [y_{(1)}, \ldots, y_{(d)}]$, $B_{p\times d} = [\beta_{(1)}, \ldots, \beta_{(d)}]$ and $d < n$. This is a multivariate regression model. We will need to refer to both rows and columns of B and use the following compact notation,

$$B = [\beta_{rc}] = [\beta_{(1)}, \ldots, \beta_{(d)}] = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_p^T \end{bmatrix},$$

and similarly for Y and X matrices. We seek a row sparse B by minimizing the following vector l_0 penalized least squares criterion.

$$J(B) = \sum_{c=1}^{d} \|y_{(c)} - X\beta_{(c)}\|^2 + h\sum_{r=1}^{p} I(\|\beta_r\| \neq 0), \quad (2)$$

where β_r, a d dimensional vector, is the r^{th} row of the B matrix and $\|\ \|$ is the Euclidean norm. $I(\|\beta_r\| \neq 0) = 1$ if $\|\beta_r\| \neq 0$ and otherwise it equals 0. The first term of the criterion determines the quality of fit and the second term introduces sparsity by penalizing the rows of B. h is the penalty parameter that determines the tradeoff between quality of fit and sparsity.

The vector l_0 criterion was used previously in another context in [22]; it removes complete rows of B in one go. It should not be confused with the scalar l_0 penalty $\sum_1^p \sum_1^d I(\beta_{rc} \neq 0)$ which only removes individual elements of B.

3. V-L0LS-CD

We now derive a cyclic descent iteration for minimizing (2). We call the algorithm V-L0LS-CD (Vector l_0 penalized Least Squares via Cyclic Descent). We obtain,

Result I: V-L0LS-CD. Given B^{k-1} the update for the u^{th} row β_u of B at the k^{th} iteration is,

$$\beta_u^k = \frac{z_u^{k-1}}{\|x_{(u)}\|^2} I(\|z_u^{k-1}\| > \|x_{(u)}\|\sqrt{h}), \quad (3)$$

where $z_u^{k-1} = E_{-u}^{k-1,T} x_{(u)}$, $E_{-u}^{k-1} = Y - X_{-u}B_{-u}^{k-1}$, X_{-u} is X with its u^{th} column removed and B_{-u}^{k-1} is B^{k-1} with its u^{th} row removed.

Proof: The least squares term of the criterion can be rewritten as $\sum_{c=1}^{d} \|y_{(c)} - X\beta_{(c)}\|^2 = \text{tr}(Y-XB)^T(Y-XB)$ $= \text{tr}(Y^T - B^T X^T)(Y^T - B^T X^T)^T$, where tr$(\cdot)$ is the trace of the matrix.

At the k^{th} iteration, decompose $k = lp + u$ where l is an integer and $1 \leq u \leq p$. Fix all the β_r's at their value at the k^{th} iteration except for β_u. Then we can rewrite the criterion as,

$$J(\beta_u) = \text{tr}(Y^T - B_{-u}^{k-1,T} X_{-u}^T - \beta_u x_{(u)}^T)(Y^T - B_{-u}^{k-1,T} X_{-u}^T - \beta_u x_{(u)}^T)^T + h\sum_{r\neq u} I(\|\beta_r^{k-1}\| \neq 0) + hI(\|\beta_u\| \neq 0),$$

$$J(\beta_u) = \text{tr}(E_{-u}^{k-1,T} E_{-u}^{k-1}) - 2\text{tr}(\beta_u x_{(u)}^T E_{-u}^{k-1}) + \|\beta_u\|^2 \|x_{(u)}\|^2 + h\sum_{r\neq u} I(\|\beta_r^{k-1}\| \neq 0) + hI(\|\beta_u\| \neq 0).$$

Set $z_u^{k-1} = E_{-u}^{k-1,T} x_{(u)}$, then $\text{tr}(\beta_u x_{(u)}^T E_{-u}^{k-1}) = z_u^{k-1,T}\beta_u$. Add and subtract $\|z_u^{k-1}\|^2/\|x_{(u)}\|^2$ from $J(\beta_u)$. Then drop terms that do not depend on β_u to get $R(\beta_u)$,

$$R(\beta_u) = \left(\frac{z_u^{k-1}}{\|x_{(u)}\|} - \|x_{(u)}\|\beta_u\right)^2 + hI(\|\beta_u\| \neq 0).$$

The minimizer of $R(\beta_u)$ delivers the cyclic descent update. $R(0) = \|z_u^{k-1}\|^2/\|x_{(u)}\|^2$, while for $\beta_u \neq 0$, $R(\beta_u)$ is minimized at $\beta_u = z_u^{k-1}/\|x_{(u)}\|^2$ giving minimized value h. Thus the minimum is at 0 if $\|z_u^{k-1}\|^2/\|x_{(u)}\|^2 \leq h$ or equivalently if $\|z_u^{k-1}\| \leq \|x_{(u)}\|\sqrt{h}$ and the result I follows.

We can further express the update as follows,

$$\frac{z_u^{k-1}}{\|x_{(u)}\|^2} = \frac{1}{\|x_{(u)}\|^2}[E_{-u}^{k-1,T} x_{(u)}], = \frac{1}{\|x_{(u)}\|^2}[E^{k-1,T} x_{(u)}] + \beta_u^{k-1},$$

where $E^{k-1} = Y - XB^{k-1}$. Thus,

$$\frac{z_u^{k-1}}{\|x_{(u)}\|^2} = \gamma_u^{k-1} + \beta_u^{k-1},$$

where $\gamma_u^{k-1} = [E^{k-1,T} x_{(u)}]/\|x_{(u)}\|^2$. We thus obtain,

Result II: The V-L0LS-CD update of result I can be re-expressed as,

$$\beta_u^k = (\gamma_u^{k-1} + \beta_u^{k-1})\, I(\|x_{(u)}\|\|\gamma_u^{k-1} + \beta_u^{k-1}\| > \sqrt{h}). \quad (4)$$

3.1. Optimality Conditions

Optimality conditions for scalar regression with scalar l_0 penalty were derived in [21]. Here we extend them to multivariate regression with vector l_0 penalty.

Result III: Optimality conditions for $J(B)$,

Define $\Gamma_0 = \{j : \|\dot{\beta}_j\| = 0\}$, $\Gamma_c = \{j : \|\dot{\beta}_j\| \neq 0\}$. Then $\dot{B}$ is a local minimum of $J(B)$ iff,

(a) $\|x_{(u)}\|\|\gamma_j\| \leq \sqrt{h}$, $j \in \Gamma_0$.

(b) $\gamma_j = 0$, $j \in \Gamma_c$.

(c) $\|x_{(u)}\|\|\dot{\beta}_j\| > \sqrt{h}$, $j \in \Gamma_c$.

Proof: The result can be established by an analysis of the fixed points of (4) and a modification of the method of [21]; details will be given elsewhere.

3.2. V-L0LS-CD algorithm

From result III it is clear that the algorithm will terminate at a local minimum. Thus proper initialization is very important especially for under-determined systems. We have tried various initialization methods and found that initializing with the solution to the l_1 penalized least squares problem B_{l_1} seems to provide the best results. Further discussion is provided in section 4.

Given X and Y, set $B^0 = B_{l_1}$. Start from $k = 1$ and increment k by one at the end of each iteration. At the k^{th} iteration decompose k and find u, find the value of β_u^k from (4) and update B. Every time p iterations get completed, check if the termination criterion is met. The algorithm can be terminated when $J(B^k) - J(B^{k+1}) \leq$ tolerance or when (a) and (b) of the optimality conditions given in section 3.1 are met.

3.3. Penalty Parameter Selection

The penalty parameter h determines the emphasis given to the two terms of the criterion (2). The importance of proper selection of h has been widely neglected in the literature. [10] sets $h = 3$ for a range of signal to noise ratios (SNR) and [17] uses an l-curve method to select the h. However the l-curve method has been heavily criticized in [23], [24]. We use the Bayesian information criterion (BIC) to select h.

4. SIMULATION

For all the simulations the data was generated as follows. The dictionary X is created by entries from a Gaussian random variable with 0 mean and unit variance. We scaled the columns of X to have unit norm $\|x_{(u)}\| = 1, u = 1, \ldots, p$. Sparsity of B for an under-determined system is $1 - k/n$, where k is the number of non zero rows of B. The locations of the non zero rows were selected from a discrete uniform distribution and the non zero rows were created by entries from a Gaussian random variable with 0 mean and unit variance. For a given X, B and SNR value the Y was generated from (1), where E contain noise vectors of zero mean and σ^2 variance. σ^2 depends on the SNR level and we assume that the noise vectors are independent from each other (correlation matrix =$\sigma^2 I$).

$$\text{SNR} = \frac{\sum_{u=1}^{d} \|X\beta_{(u)}\|^2}{n \times d \times \sigma^2}.$$

A preliminary set of simulations is done to get h by BIC. This h is then used in a second set of simulations to study the algorithm performance. The selected h is kept fixed at each iteration within an algorithm unlike in [17].

The performance of an algorithm can be measured by the Parameter mean squared error (MSE),

$$\text{Parameter MSE} = E\left(\frac{\sum_{u=1}^{d} \|\hat{\beta}_{(u)} - \beta_{(u)}\|^2}{\sum_{u=1}^{d} \|\beta_{(u)}\|^2}\right),$$

where $\hat{\beta}_{(u)}$ are the columns of the estimate $\hat{B}$ and $\beta_{(u)}$ are the columns of the original B matrix. $E(\cdot)$ denote the expected value. The quality of the estimate can also be measured by how well the estimate recovers the original model. Define $\Gamma_0 = \{u : \|\beta_u\| = 0\}$, $\Gamma_c = \{u : \|\beta_u\| \neq 0\}$, $\hat{\Gamma}_0 = \{u : \|\hat{\beta}_u\| = 0\}$ and $\hat{\Gamma}_c = \{u : \|\hat{\beta}_u\| \neq 0\}$. Then we can define true positive (TP)$= |\Gamma_c \bigcap \hat{\Gamma}_c|$, false negative (FN) $= |\Gamma_c \bigcap \hat{\Gamma}_0|$, false positive (FP) $= |\Gamma_0 \bigcap \hat{\Gamma}_c|$ and true negative (TN) $= |\Gamma_0 \bigcap \hat{\Gamma}_0|$, where $|\ |$ represent the cardinality of the set. Now we can define true positive rate (TPR$=$ TP$/$(TP $+$ FN)) and false positive rate (FPR$=$ FP$/$(FP $+$ TN)) which are important performance indicators to measure how well the estimates select the correct model.

We will compare V-L0LS-CD with vector l_1 penalized least squares [16], regularized M-FOCUSS [17], JLZA [10](we use tuning parameter settings recommended in [10]) and SOMP [13]. Vector l_1 penalized least squares is introduced in [16], but is solved by second order cone programming; instead we use cyclic descent(V-L1LS-CD). Since the criterion is convex, both algorithms will produce the same answer. As stated in [17], we set $p = 0.8$ and at the end of the algorithm we perform orthogonal projection of Y on to the atoms selected by the algorithm. [17] perform hard thresholding of the estimates of regularized M-FOCUSS so that the sparsity of the estimates would equal that of the original B matrix. We omitted this step as the sparsity of the original B matrix is generally unknown. We initialize V-L0LS-CD with all zeros $B^0 = 0$ as well as with the estimate of V-L1LS-CD $B^0 = B_{l_1}$ to show the importance of initialization.

We use dimensions similar to [17]; $n = 20$, $p = 30$. X is kept it fixed through out the simulation.

First we investigate the variation of TPR, FPR and Parameter MSE with sparsity. We set $d = 3$ and SNR$=$ 10 and vary k from 2 to 10. For each sparsity level we generated 50 B matrices and using each B matrix we generated 100 Y matrices. Results are given in figure 1.

V-L1LS-CD has the highest TPR, however it also has the highest FPR. This means that V-L1LS-CD produces estimates with very low sparsity and is thus undesirable. V-L0LS-CD with $B^0 = B_{l_1}$ has the next highest TPR while maintaining the lowest FPR. Furthermore V-L0LS-CD with $B^0 = B_{l_1}$ has the lowest Parameter MSE specially towards the lower sparsity levels. Thus in this example V-L0LS-CD with $B^0 = B_{l_1}$ is superior to the others.

Secondly we investigated the variation of TPR, FPR and Parameter MSE with SNR. We fixed $d = 2$, $k = 7$ and varied SNR from 30 to 3. Results are given in figure 2.

Similar to the earlier example V-L1LS-CD has very high FPR and V-L0LS-CD with $B^0 = B_{l_1}$ has the lowest FPR. Furthermore V-L0LS-CD with $B^0 = B_{l_1}$ has the lowest Parameter MSE.

From both these examples it is clear that when considered individually V-L1LS-CD produces very low sparsity results with very high FPR and V-L0LS-CD with $B^0 = 0$ produces

results with very low TPR. However when V-L0LS-CD is initialized with V-L1LS-CD estimate it produces the best results.

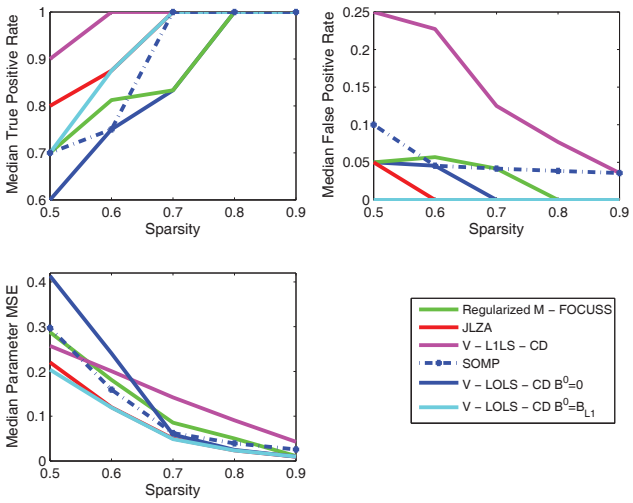

Fig. 1. Variation with sparsity.

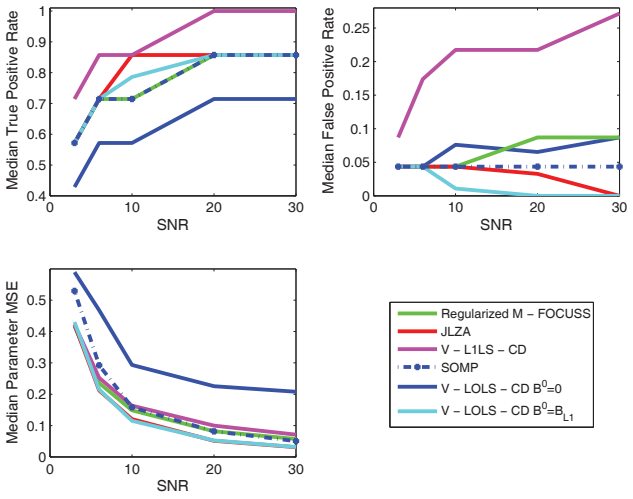

Fig. 2. Variation with SNR.

5. CONCLUSION

In this paper a cyclic descent based algorithm(V-L0LS-CD) was proposed to minimize the vector l_0 penalized least squares criterion. We have presented the optimality conditions of the algorithm and discussed the proper selection of the penalty parameter. The simulation results show that V-L0LS-CD initialized with the estimate of V-L1LS-CD produces superior results in terms of TPR, FPR and Parameter MSE when compared with existing algorithms.

6. ACKNOWLEDGMENT

The authors like to thank the School of Computer Science and Engineering at UNSW for providing access to a fast computer cluster to run the simulations.

7. REFERENCES

[1] A. J. Miller, *Subset Selection in Regression*, London: Chapman and Hall, 2002.

[2] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, pp. 2231 – 2242, 2004.

[3] S. Alliney and S. A. Ruzinsky, "An algorithm for the minimization of mixed l_1 and l_2 norms, with application to bayesian estimation," *IEEE Trans. Sig. Proc.*, vol. 42, pp. 618 – 627, 1994.

[4] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Jl. Roy. Stat. Soc. B*, vol. 58, pp. 267–288, 1996.

[5] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans.Sig.Proc.*, vol. 47, pp. 187–200, 1999.

[6] G. H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed l^0 norm," *IEEE Trans. Sig. Proc.*, vol. 57, pp. 289–301, 2009.

[7] G. A. F. Seber, *Multivariate Observations*, J. Wiley, 1984.

[8] T. W. Anderson, *An introduction to multivariate statistical analysis*, New York: Wiley, 1958.

[9] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with focuss: a recursive weighted minimum norm algorithm," *Electroen. Clin. Neuro.*, vol. 95, pp. 231–251, 1995.

[10] M. Hyder and K. Mahata, "Direction-of-arrival estimation using a mixed $l_{2,0}$ norm approximation," *IEEE Trans. Sig. Proc.*, vol. 58, pp. 4646–4655, 2010.

[11] I. J. Fevrier, S. B. Gelfand, and M. P. Fitz, "Reduced complexity decision feedback equalization for multipath channels with large delay spreads," *IEEE Trans. Commun.*, vol. 47, pp. 927–937, 1999.

[12] B. D. Jeffs, "Sparse inverse solution methods for signal and image processing applications," *ICASSP, Seattle, Washington, U.S.A.*, vol. 3, pp. 1885–1888, 1998.

[13] J. A. Tropp, A. C. Gilbert, , and M. J. Strauss, "Algorithms for simultaneous sparse approximation. part i: Greedy pursuit," *Sig. Proc.*, vol. 86, pp. 572–588, 2006.

[14] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms," *Jl. of Fourier Anal. Applic.*, vol. 14, pp. 655–687, 2008.

[15] J. A. Tropp, "Algorithms for simultaneous sparse approximation. part ii: Convex relaxation," *Sig. Proc.*, vol. 86, pp. 589–602, 2006.

[16] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Sig. Proc.*, vol. 53, pp. 3010–3022, 2005.

[17] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Sig. Proc.*, vol. 53, pp. 2477–2488, 2005.

[18] M. Mishali and Y. C. Eldar, "Reduce and boost: Recovering arbitrary sets of jointly sparse vectors," *IEEE Trans. Sig. Proc.*, vol. 56, pp. 4692–4702, 2008.

[19] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Sig. Proc.*, vol. 54, pp. 4634–4643, 2006.

[20] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inf. Theory*, vol. 56, pp. 505–519, 2010.

[21] T. Blumensath and M.E. Davies, "Iterative thresholding for sparse approximation," *Jl. of Fourier Anal. Applic.*, vol. 14, pp. 629–654, 2008.

[22] M.O. Ulfarsson and V. Solo, "Sparse variable noisy pca using l_0 penalty," ICASSP, Dallas, Texas, U.S.A., 2010, pp. 3950 – 3953.

[23] M. Hanke, "Limitations of the l-curve method in ill-posed problems," *BIT Numerical Mathematics*, vol. 36, pp. 287–301, 1996.

[24] C. R. Vogel, "Non-convergence of the l-curve regularization parameter selection method," *Inverse Probl.*, vol. 12, pp. 535–547, 1996.