A MODIFIED CLEAN ALGORITHM DOES L_1-DENOISING.

Victor Solo

School of Electrical Engineering and Telecommunications
University of New South Wales
Sydney, AUSTRALIA
email: v.solo@unsw.edu.au

ABSTRACT

The CLEAN algorithm is one of the best known signal processing algorithms in (radio-)astronomy. It is essentially a deconvolution procedure and is used e.g. to reconstruct a sparse (star) brightness distribution from noisy ('dirty' - hence 'cleaning') observed data. In this paper we make a connection for the first time between CLEAN and l_1-denoising. We show CLEAN does a crude version of l_1-denoising and develop a modified algorithm with much improved behaviour.

Index Terms— l_1 denoising, sparse, CLEAN, estimation.

1. INTRODUCTION

The CLEAN algorithm is widely used in Astronomy as a means of reconstructing sparse spatial point sources from noisy data. The algorithm was developed by [1] but later [2] traced it back to [3],[4],[5].

While CLEAN is usually presented in an Astronomical applications contex, as [2] made clear it is just an (unusual) iterative algorithm for solving a linear regression problem.

In practice the algorithm is not run to convergence but stopped with an ad-hoc criterion. And it is the analysis of the iteration together with especially its stopping criterion that forms the point of departure for our work.

l_1-denoising has gained a lot of interest as a signal estimation procedure in recent years. Particularly because of its ability to produce sparse solutions. l_1 penalised optimization approaches to signal estimation seem to have a number of origins. They were developed in image processing ; see[6] for references, a summary and important convergence analysis, and independently [7]. But in these image processing applications it is an image gradient that is penalised. When the penalty is used on signal amplitude rather than signal gradient the origins are different. Important earlier work is [8]. The l_1 penalised least squares regression problem of interest here was apparently first formulated in [9] and later independently by [10] where it was given the name LASSO. To recognize the priority of [9] we henceforth refer to it as LART. Although the l_1 penalized least squares regression problem can be solved by linear programming, iterative procedures remain of interest and [11] developed an important algorithm building on work of [12]. For inverse problems work see [13] and its references.

In section 2 we set up the least squares regression problem. In section 3 we recap the CLEAN algorithm and its convergence analysis. In section 4 we develop the connexion to l_1-denoising. This leads us to modify the CLEAN algorithm to obtain an iterative l_1-denoiser.. An example is given in section 5 and conclusions in section 6.

2. REGRESSION

As indicated in the introduction we will view CLEAN simply as an unusual iteration for solving a least squares regression problem . Given measurements $y_i, x_i, i = 1, \cdots, n$ on a dependent variable y and predictors or regressors $x_i^T = (x_{1,i}, \cdots, x_{p,i})$, the problem is to fit a linear regression relation by least squares , $\hat{c}, \hat{\beta} = arg.\,^{min}_{c,\beta}\, \Sigma_1^n (y_i - c - x_i^T \beta)^2$ where c is an intercept. In matrix form this is

$$\hat{c}, \hat{\beta} = arg\,^{min}_{c,\beta}\ \parallel y - c1_n - X\beta \parallel^2$$

and $y = (y_1, \cdots, y_n)^T$; $1_n = (1, \cdots, 1)^T$ is an n-vector; $X^T = (x_1, \cdots, x_n)$. The columns of X which contain all the observations on each variable will be denoted as $x_{(u)}, u = 1, \cdots, n$ so $X = (x_{(1)}, \cdots, x_{(p)})$.

It is important for the subsequent development that we first centre the variables about their means. This removes c from the problem. We continue to denote the centred columns by $x_{(u)}$ and the data by y. The problem is now

$\hat{\beta} = arg.\,^{min}_{\beta}\, E(\beta) : E(\beta) = \frac{1}{2} \parallel y - X\beta \parallel^2$

The mean vector $\mu = X\beta$, can be written $\mu = \Sigma_1^p x_{(u)} \beta_u$. We now require the columns be scaled to unit length. So with $b_u = \parallel x_{(u)} \parallel \beta_u, u = 1, \cdots, p$,

$$\mu = \Sigma_1^p x_{(u)} \beta_u = \Sigma_1^p \frac{x_{(u)}}{\parallel x_{(u)} \parallel} b_u$$

The utility of column scaling in improving the numerical conditioning of X has long been understood in numerical analy-

sis [14] and statistics [15]. For simplicity we will now denote the scaled columns as $x_{(u)}$ so $\mu = \Sigma_1^p x_{(u)} b_u$.

If $X^T X$ has full rank, the least squares problem has solution $b_{ols} = (X^T X)^{-1} X^T Y$. However if either this is not the case or X is poorly conditioned as in ill-conditioned inverse problems, this solution is not available. Also in large dimensional problems the computation may be prohibitive. For these reasons and others to be clear, iterative solutions continue to be of interest. In the sequel if $X^T X$ is not of full rank we take b_{OLS} to denote a solution of the least squares equations $X^T(y - X b_{OLS}) = 0$.

There are many iterative procedures for solving the least squares regression problem such as Gauss-Seidel, Landweber etc [16]. But CLEAN is very different from all these.

3. CLEAN & RELATED ALGORITHMS

We now proceed to describe and review the CLEAN algorithm. The algorithm is initiated at $b^{(0)} = 0$.. Given the iterate $b^{(k)}$ we proceed as follows. We first calculate the current residual or error vector, $e^{(k)} = y - X b^{(k)}$. We also introduce the negative of the E-gradient

$$\gamma^{(k)} = X^T e^{(k)} \, , \equiv \gamma_u^{(k)} = x_{(u)}^T e^{(k)}, 1 \leq u \leq p$$

Since the $x_{(u)}$ are centred and scaled, $\gamma_u^{(k)}$ is the covariance between the uth variable and the current residual.

The idea is to update the regression vector by changing just one component. So we look for an update

$$b^{(k+1)} = b^{(k)} + \alpha \rho \delta_u \tag{3.1}$$

where δ_u is a vector of 0s but with a 1 in position u. And the step ρ and the index u are to be chosen. Also α is a gain factor which we take to be 1 for the moment.

3.1. CLEAN

We aim to choose ρ, u to maximize the reduction in error sum of squares . The residual after the update is

$$\begin{aligned} e^{(k+1)} &= y - X b^{(k+1)} \\ &= y - X(b^{(k)} + \rho \delta_u) = e^{(k)} - \rho x_{(u)} \end{aligned}$$

And the error sum of squares after the update is

$$\| e^{(k+1)} \|^2 = \| e^{(k)} \|^2 - 2\rho x_{(u)}^T e^{(k)} + \rho^2 \tag{3.2}$$

since $\| x_{(u)} \|^2 = 1$. We now choose ρ, u as,

$$\hat{u}, \hat{\rho} = arg \min_u \min_\rho \| e^{(k+1)} \|^2$$

For given u, $\| e^{(k+1)} \|^2$ is clearly minimized at $\hat{\rho}_u = x_{(u)}^T e^{(k)} = \gamma_u^{(k)}$ giving a minimized error sum of squares of

$$\| e_u^{(k+1)} \|^2 = \| e^{(k)} \|^2 - (\gamma_u^{(k)})^2$$

We now choose u to minimize this leading to

$$\hat{u}_k = \hat{u} = arg \max_u |\gamma_u^{(k)}| = arg \max_u |x_{(u)}^T e^{(k)}| \tag{3.3}$$

So $\hat{u}$ is the index corresponding to the variable whose gradient or covariance is largest. This then yields a step

$$\hat{\rho} = x_{\hat{u}}^T e^{(k)} = \gamma_{\hat{u}}^{(k)} \tag{3.4}$$

and an error sum of squares

$$\| e^{(k+1)} \|^2 = \| e^{(k)} \|^2 - (\gamma_{\hat{u}}^{(k)})^2 \tag{3.5}$$

Note that if we now allow $\alpha \neq 1$ in the update (3.1) but still use (3.4) then the error sum of squares update (3.5) becomes

$$\| e^{(k+1)} \|^2 = \| e^{(k)} \|^2 - \alpha(2 - \alpha)(\gamma_{\hat{u}}^{(k)})^2 \tag{3.6}$$

This now yields the convergence result of [4], [5],[2].

Theorem 1. For the algorithm (3.1,3.3,3.4) with $0 < \alpha < 2$ then $b^{(k)} \to b_{OLS}$ as $k \to \infty$.

Proof. From 3.6, so long as $0 < \alpha < 2$, the error sum of squares continues to reduce until $\gamma_{\hat{u}}^{(k)} = 0$ whereupon by definition of $\hat{u}$ all gradients vanish. Thus the converged value obeys $X^T(y - X b^\infty) = 0 \Rightarrow b^\infty = b_{OLS}$.

Below we use $\alpha = 1$ since it gives the fastest convergence.

3.2. RELATED ALGORITHMS

Although no awareness of CLEAN is shown in [11] an algorithm is introduced, called forward stagewise regression, in which $\gamma_{\hat{u}}^{(k)}$ is replaced by its sign. We call it sign-CLEAN (sCLEAN). Next is forward selection regression, well known in variable selection in regression. At each iteration one adds the variable showing a maximum correlation; but it is a different (though related) correlation to that used by CLEAN. Also the regression vector update is totally different. When a new variable is added, the whole regression is redone and the new regression vector is the least squares estimator. Other important algorithms that differ from CLEAN are the LARS algorithm [11] and the Landweber algorithm [17] (who calls it the shooting algorithm) which was rediscovered by [18].

4. L_1-DENOISING

Consider the following penalized least squares problem ,
$\min_b J(b) : J(b) = \frac{1}{2} \| y - Xb \|^2 + h\Sigma_1^p |b_u|$
The criterion is easily seen to be convex in b, so that there is a unique minimum, denoted b_*. The derivative of the penalty (the J-gradient) is

$$\frac{\partial J}{\partial b_u} = -x_{(u)}^T (y - Xb) + h sgn(b_u) = -\gamma_u + h sgn(b_u)$$

is discontinuous at 0 for each b_u. This means the solution involves exact zeroing of some components of b,[9],[10]. Thus

sparse solutions can be obtained. The amount of sparseness is controlled by the penalty weight h. The empirical choice of h is an important issue that will be discussed elsewhere. One approach is given in [11].

The first order optimality conditions are given as follows [9]. Let $Z = \{u : b_u = 0\}, Z^c = \{u : b_u \neq 0\}$. Then

$$\gamma_u - hsgn(b_u) = 0, u \in Z^c \quad (4.7)$$

$$|\gamma_u| \leq h, u \in Z \quad (4.8)$$

Note that it follows that, for all $u, |\gamma_u| \leq h$.

[9] develop an iterative procedure for minimizing $J(b)$ but it requires a messy monitoring procedure and is quite different from CLEAN.

4.1. CLEAN Revisited

In practice CLEAN is not iterated to convergence. Rather the E-gradient is monitored and the iteration is terminated when $|\gamma_{\hat{u}}^{(k)}| \leq \lambda$ where $\lambda > 0$ is a user chosen threshold. The similarity of this stopping condition to the first order optimality conditions for LART is quite striking and it is this that first caught the author's attention. We now have.

Theorem 2. CLEAN with the rule: stop the iteration at the first k for which $|\gamma_{\hat{u}}^{(k)}| \leq 2h$, iteratively reduces $J(b)$.

$$J(b^{(k+1)}) < J(^{(k)}), \text{ while } |\gamma_{\hat{u}}^{(k)}| > 2h$$

. *Proof.* We have from (3.5,3.4)

$$J(b^{(k+1)}) = \frac{1}{2} \parallel e^{(k)} \parallel^2 - \frac{1}{2}(\gamma_{\hat{u}}^{(k)})^2 + h\Sigma_1^p |b_u^{(k+1)}|$$

But from (3.1,3.4), $b_u^{(k+1)} = b_u^{(k)}, u \neq \hat{u}$ while

$$\begin{aligned} |b_{\hat{u}}^{(k+1)}| &= |b_{\hat{u}}^{(k)} + \gamma_{\hat{u}}^{(k)}| \leq |b_{\hat{u}}^{(k)}| + |\gamma_{\hat{u}}^{(k)}| \\ \Rightarrow J(b^{(k+1)}) &\leq J(b^{(k)}) - \frac{1}{2}(\gamma_{\hat{u}}^{(k)})^2 + h|\gamma_{\hat{u}}^{(k)}| \\ &< J(b^{(k)}), \text{ if } |\gamma_{\hat{u}}^{(k)}| > 2h. \end{aligned}$$

This is a most remarkable and encouraging result and establishes the formal connexion between CLEAN and l_1-denoising. But CLEAN falls short of minimizing $J(b)$ because it requires we stop when $|\gamma_{\hat{u}}^{(k)}| \leq 2h$ whereas we need to continue reducing $J(b)$ until $|\gamma_{\hat{u}}^{(k)}| \leq h$.

A similar result holds for sign-CLEAN with $\alpha = 2h$. This choice of α however means very slow convergence. And again we have the problem of stopping early.

4.2. STM-CLEAN

Fortunately a judicious combination of CLEAN and sCLEAN overcomes this. The regression vector update uses the negative E-gradient. A natural idea is to use the negative J-gradient instead. From the condition (4.7) we note $sgn(\gamma_u) = sgn(b_u), b_u \neq 0$. So we modify the negative J-gradient from $\gamma_u - hsgn(b_u)$ to $\gamma_u - hsgn(\gamma_u) = sgn(\gamma_u)(|\gamma_u| - h)$.

This leads to a new algorithm:

STM-CLEAN = CLEAN with a soft thresholding modification. The update is $b^{(k+1)} = b^{(k)} + \hat{\rho}\delta_{\hat{u}}$ where,

$$\hat{\rho} = sgn(\gamma_{\hat{u}}^{(k)})\hat{R}_k, \hat{R}_k = |\gamma_{\hat{u}}^{(k)}| - h$$

Note then that, $\parallel b^{(k+1)} - b^{(k)} \parallel_\infty = |\hat{R}_k|$. We now introduce the limit set $L = \{b : \max_u |\gamma_u(b)| = h\}$. Note that the optimal value $b_* \in L$. And any limit point of STM-CLEAN must lie in L. Continuing, we can take $\hat{R}_o > 0$ since otherwise the start value is already in L.

We now have:

Theorem 3. STM-CLEAN has the following properties:
(i) After each iteration there is at least one index, namely, $\hat{u} = \hat{u}_k$ for which $|\gamma_u^{(k+1)}| = h$.
(ii) $\hat{R}_k \geq 0, k \geq 0$
(iii) $\hat{R}_k \to 0$ as $k \to \infty$.
(iv) $b^{(k)}$ converges to a compact connected subset of L.
Remark. We are not able to show $b^{(k)} \to b_*$ but simulations below suggest it does.

Proof. The E-gradient update for index $\hat{u} = \hat{u}_k$ is $\gamma_{\hat{u}}^{(k+1)} = \gamma_{\hat{u}}^{(k)} - (\gamma_{\hat{u}}^{(k)} - hsgn(\gamma_{\hat{u}}^{(k)})) = hsgn(\gamma_{\hat{u}}^{(k)})$. Which yields $|\gamma_{\hat{u}}^{(k+1)}| = h$ and (i) is established. Now due to (i) at iteration $k+2$ we will have $\hat{R}_{k+1} \geq 0$ and this holds for $k \geq 1$ and hence for $k \geq -1$ so (ii) holds. Proceeding much as before, (3.2) gives,

$$\begin{aligned} \parallel e^{(k+1)} \parallel^2 &= \parallel e^{(k)} \parallel^2 - 2\hat{\rho}\gamma_{\hat{u}}^{(k)} + \hat{\rho}^2 \\ &= \parallel e^{(k)} \parallel^2 - 2\hat{R}_k|\gamma_{\hat{u}}^{(k)}| + \hat{R}_k^2 \end{aligned}$$

Repeating the argument in Theorem 2 we find, since $\hat{R}_k \geq 0$,

$$\begin{aligned} J(b^{(k+1)}) - J(b^{(k)}) &\leq -|\gamma_{\hat{u}}^{(k)}|\hat{R}_k + \frac{\hat{R}_k^2}{2} + h|\hat{R}_k| \\ \leq -(\hat{R}_k + h)\hat{R}_k + \frac{\hat{R}_k^2}{2} + h\hat{R}_k &\leq -\frac{1}{2}\hat{R}_k^2 \end{aligned}$$

Now $J(b)$ is lower bounded by $J(b_*) > 0$ and so $J(b^{(k)})$ is a lower bounded non-increasing sequence and so must have a finite limit say J_∞. Now summing up the inequality we find $\frac{1}{2}\Sigma_0^\infty \hat{R}_k^2 + J_\infty \leq J(b^{(0)}) < \infty$. Thus $\hat{R}_k \to 0$ as $k \to \infty$. Thus $\parallel b^{(k+1)} - b^{(k)} ||_\infty \to 0$. Then by Ostrowski's theorem [19],$b^{(k)}$ converges to a compact connected set which must be a subset of L.

5. SIMULATIONS

We simulated data from a regression model $y = Xb + \epsilon$ where the entries of X, ϵ are independent identically distributed Gaussians with zero mean and unit variance. X was generated only once and then column scaled; ϵ changed with

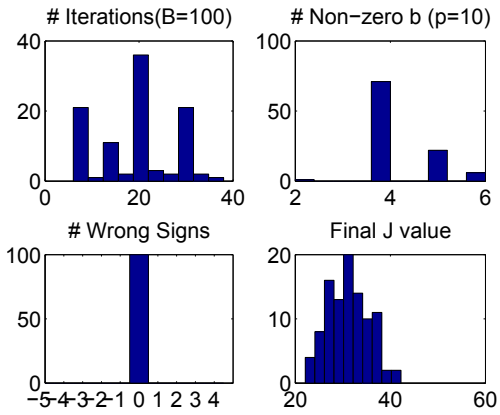

Fig. 1. STM-CLEAN Simulation Histograms

each simulation. The regession vector was also generated as a Gaussian random vector and then thresholded and rounded leaving 4 non-zero coefficients. We used n=200,p=10,h=2 and B=100 simulations.

Fig.1. shows for STM-CLEAN histograms of # of iterations, final J-values, # non-zero coefficients, # wrong signs (i.e. when sign$\gamma_u \neq sign(b_u)$) -there are none - at convergence. A similar set of histograms (done on the same data) for CLEAN shows not surprisingly, mostly wrong signs, much higher final J values and many fewer iterations. Plots not shown indicate that after an initial rapid reduction most of the iterations appear to be used for making small adjustments.

6. CONCLUSIONS

In this paper we have exhibited for the first time a relation between the celebrated CLEAN algorithm of astronomy and l_1-denoising. This has led us to a new algorithm STM-CLEAN which comes closer to solving the l_1-denoising problem.

7. REFERENCES

[1] J.A. Hogbom, "Aperture synthesis with a nonregular distribution of interferometer baselines," *Astron. Astrophys. Suppl.*, vol. 15, pp. 417–426, 1974.

[2] UJ Schwarz, "Mathematical-statistical description of the iterativve beam removing technique (method clean)," *Astron. Astrophys*, vol. 65, pp. 345–356, 1978.

[3] RV Southwell, "Stress-calculation in frameworks by the method of systematic relaxation of constraints: I and ii," *Proc. Roy. Soc. A,*, vol. 151, pp. 56–95, 1935.

[4] G Temple, "The general theory of relaxation methods applied to linear systems," *Proc. Roy. Soc. A*, vol. 169, pp. 476–500, 1939.

[5] GE Forsythe and WR Wasow, *Finite Difference Methods for Partial Differential Equations*, J Wiley, New York, 1960.

[6] P Charbonnier, L Blanc-Feraud, G Aubert, and M Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE TRans Image Processing*, vol. 6, pp. 298–311, 1997.

[7] L.I. Rudin, S.Osher, and E.Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, pp. 259–268, 1992.

[8] DL Donoho, IM Johnstone, JC Hoch, and AS Stern, "Maximum entropy and the nearly black object," *Jl. Royal. Stat. Soc. Ser. B*, vol. 54, pp. 41–81, 1992.

[9] S Alliney and SA Ruzinsky, "An algorithm for the minimization of mixed l_l and l_2 norms with application to bayesian estimation," *IEEE Transactions on Signal Processing*, vol. 42, pp. 618–627, 1994.

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

[11] B Efron, T Hastie, I Johnstone, and R Tibshirani, "Least angle regression," *Ann. Stat.*, vol. 32, pp. 407–499, 2004.

[12] MR Osborne, B Presnell, and BA Turlach, "A new approach to variable selection in least squares problems," *IMA Jl Num Anal*, vol. 20, pp. 389–404, 2000.

[13] SF Cotter, BD Rao, K Engan, and K Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Transactions on Signal Processing*, vol. 53, pp. 2477–2488, 2005.

[14] GW Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1974.

[15] RE Welsch DA Belsey, E Kuh, *Regression Diagnostics*, John Wiley, New York, 1980.

[16] PC Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM, Philadelphia, 1998.

[17] WJ Fu, "The bridge versus the lasso," *Jl Comp. Graph. Stat.*, vol. 7, pp. 397–416, 1998.

[18] I Daubechies, M Defrise, and C De-Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm Pure Appl Math*, vol. LVII, pp. 1413–1457, 2004.

[19] AM Ostrowski, *Solution of equations in Euclidean and Banach spaces*, Academic Press, New York, 1973.