REGULARIZED FOCUSS FOR SUBSET SELECTION IN NOISE

Kjersti Engan

Electrical and Comp. Eng. Dept.
Høgskolen i Stavanger
P.O.Box 2557 Ullandhaug
4091 Stavanger, NORWAY
e-mail Kjersti.Engan@tn.his.no

Bhaskar D. Rao and K. Kreutz-Delgado

Electrical and Comp. Eng. Dept.
University of California, San Diego
La Jolla, CA 92093-0407
e-mail: {brao,kreutz}@ece.ucsd.edu

ABSTRACT

This paper considers subset selection in the presence of noise via algorithms that minimize diversity measures. This leads to iterative procedures like regularized FOCUSS in which each iteration involves the solution to a regularized least squares problem. Several different methods for choosing the regularization parameter such as the discrepancy principle and the L-curve technique are evaluated for this purpose and compared. To overcome some of the limitations in existing methods, a modified L-curve approach is proposed. Experiments using regularized FOCUSS with the different methods for choosing the regularization parameter are conducted and compared with a sequential subset selection method, the Orthogonal Matching Pursuit (OMP) method. Results show that a modified L-curve approach works well for finding the regularization parameter in this application.

1. INTRODUCTION

Subset selection problems arise in many applications such as basis selection in signal representation [1], Magnetoencephalography [2], among others. The underlying mathematical problem is one of solving an underdetermined linear system of equations with the requirement that the solution be sparse, i.e. solving the linear inverse problem

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{1}$$

where $\dim(\mathbf{A}) = m \times n$ with $m \leq n$. The requirement that x be sparse, i.e. have only a few nonzero entries, prohibits the use of the minimum 2-norm solution and leads to the problem of subset selection. Several interesting algorithms have been developed to address this need. Of interest in this paper is a method called FOCUSS, for **FOC**al **U**nderdetermined **S**ystem **S**olver [3], that minimizes the $\ell_{(p\leq 1)}$ diversity measure given by [4]

$$E^{(p)}(\mathbf{x}) = \mathrm{sgn}(p) \sum_{i=1}^{n} |\mathbf{x}[i]|^p, \quad p \leq 1. \tag{2}$$

The research of B. Rao and K. Kreutz-Delgado was partially supported by the National Science Foundation Grant No. CCR-9902961.

The FOCUSS algorithm gives an exact solution to (1). In some applications, such as using the model for compression [5], it is desirable to allow for an error to achieve a solution that is very sparse, i.e. $\mathbf{b} = \mathbf{A}\mathbf{x} + \mathbf{v}$. In addition, the data measurements can be noisy or there are modeling errors making an error in the reconstruction meaningful.

Using a Bayesian formulation, methods based on minimizing diversity measures can be adapted to deal with noise. This approach leads to the minimization of a modified function of the type

$$J(x) = \|Ax - b\|^2 + \beta E^{(p)}(\mathbf{x}). \tag{3}$$

Using the factored gradient approach [4], a variation of the FOCUSS algorithm can be developed that can be used in the presence of noise [6]. The algorithm consists of the following iterations: ($p \leq 1$ and is selected by user)

$$\mathbf{W}_{k+1} = \mathrm{diag}(|\mathbf{x}_k[i]|^{1-\frac{p}{2}}) \tag{4}$$

$$\mathbf{q}_{k+1} = \arg\min_{\mathbf{q}} \|\mathbf{A}\mathbf{W}_{k+1}\mathbf{q} - \mathbf{b}\|^2 + \lambda\|\mathbf{q}\|^2 \tag{5}$$

$$\mathbf{x}_{k+1} = \mathbf{W}_{k+1}\mathbf{q}_{k+1}, \tag{6}$$

where (5) is a regularization problem and λ is the regularization parameter. The quality of the sparse solution obtained via the regularized version of FOCUSS is governed by the choice of λ. No systematic and robust procedure has been presented for choosing λ. This paper concentrates on this problem.

2. REGULARIZATION PARAMETER

Determining a proper value for λ is an important problem and has implications to other subset selection methods as well as to other regularization problems. Sparsity adds an interesting twist to this classical problem, and in the subset selection context, there appears to be no practical reason to limit the choice of λ to a fixed value for all the iterations. A value that is dependent on the iteration offers more flexibility and may be more appropriate. With this in mind, we suggest and study three approaches motivated by three different scenarios. The first approach

is motivated by the desire to ensure a certain quality of representation and exploits the availability of some information on the perturbations. The second by the need to ensure a certain degree of sparsity on the solution as would be required in applications like compression. The third by the desire to produce stable sparse solutions without the need for much prior information.

2.1. Discrepancy principle

In this approach, the goal is to find a sparse solution that assures a certain quality in the nature of the representation, i.e. $\|\mathbf{Ax} - \mathbf{b}\| \leq \epsilon$. This is called the *discrepancy principle* [7]. Algorithmically this reduces to solving the optimization problem

$$\min_{\mathbf{x}} E^{(p)}(\mathbf{x}) \text{ subject to } \|\mathbf{Ax} - \mathbf{b}\| \leq \epsilon.$$

The formulations avoids the need to find a regularization parameter explicitly. Viewed from a Bayesian perspective, (3), the approach amounts to a particular choice of β and hence of a regularization parameter implicitly. Assuming that the inequality constraint is active, which is usually true, and following the factored gradient approach, an iterative algorithm can be derived which at each iteration computes $\mathbf{x}_{k+1} = \mathbf{W}_{k+1}\mathbf{q}_{k+1}$, where

$$\mathbf{q}_{k+1} = \arg\min_{\mathbf{q}} \|\mathbf{q}\|^2 \text{ subject to } \|\mathbf{AW_{k+1}q} - \mathbf{b}\| \leq \epsilon.$$

An algorithm for computing $\mathbf{q}_{k+1}$ can be readily derived [9].

2.2. Sparsity Criteria

Another option is to choose λ so that the solution produced has a predetermined number of nonzero entries r. Note that upon convergence the rank of $\mathbf{AW}_{k+1}$ is equal to r, i.e. $\lim_{k\to\infty} \text{rank}(\mathbf{AW}_{k+1}) = r$. So a desirable approach would be to use a sequence λ_k to satisfy this limiting rank property, while providing the best fit possible. A reliable procedure for doing this is not yet available. One practical approach, which is used here, is to use a sequential basis selection method like the Orthogonal Matching Pursuit (OMP) to select r columns [10], and to determine a value for the error ϵ in the representation. This ϵ can be the basis of FOCUSS along the lines suggested in section 2.1. In the end, one can either use the subset selected with FOCUSS or the OMP solution, whichever is better.

2.3. Modified L-Curve Criteria

In this approach, the regularizing parameter is found by striking a compromise between minimizing the norm of the solution vector, $\|\mathbf{q}\|^2$, versus the error in the representation, $\|\mathbf{AW}_{k+1}\mathbf{q} - \mathbf{b}\|^2$. In this context, this choice also translates into controlling the sparse nature of the solution, so that a trade off between quality of fit and sparsity is done. The use of such an approach in the context of FOCUSS was first suggested in [3]. The L-curve was introduced by Hansen in [11] as a method for finding the parameter λ in the regularization problem: $\min_{\mathbf{x}}\{\|\mathbf{Ax} - \mathbf{b}\|^2 + \lambda\|\mathbf{x}\|^2\}$, and this can easily be translated to the regularization problem of (5) which we want to solve.

If λ is varied from 0 to ∞, $\|\mathbf{q}\|^2$, a measure of sparsity, decreases monotonically from $\|(\mathbf{AW}_0)^+\mathbf{b}\|^2$ to zero and $\|\mathbf{AW}_{k+1}\mathbf{q} - \mathbf{b}\|^2$, a measure of the approximation error, increases monotonically. The theory of the L-curve poses that a plot of $\|\mathbf{q}\|^2$ versus $\|\mathbf{AW}_{k+1}\mathbf{q} - \mathbf{b}\|^2$ for different λ will be shaped as an L, and that a good λ is the one corresponding to the corner in the L. Furthermore it is suggested [11, 7] that the corner of the L-shaped curve can be found by finding the maximum curvature. Though the plot of $\|\mathbf{q}\|^2$ versus $\|\mathbf{AW}_{k+1}\mathbf{q} - \mathbf{b}\|^2$ can be shown to be convex [7], and the maximum curvature can be at a trade off point between sparsity and accuracy, this is usually carried on a log-log scale. In our experiments we have found that the L-curve approach using a log-log scale often fails because the data will *not* produce a well defined L-curve in each iteration of the FOCUSS algorithm.

To improve the robustness, we propose a method using a combination of the discrepancy principle and the L-curve method, linear scale. We call this the *modified L-curve method*. In the proposed modified L-curve method, it is assumed that there is some knowledge about the variance of the noise, or alternatively something about the target SNR after computing an approximation. From this knowledge an upper and an lower target on the residual norm, $\epsilon^2 = \|\mathbf{Ax} - \mathbf{b}\|^2$ can be made. Then for every iteration in FOCUSS the upper and lower target on ϵ^2 is used to find an upper and a lower limit for λ, $\{\lambda_{min}, \lambda_{max}\}$. The λ corresponding to the maximum curvature in the linear scale, λ_c, is also calculated in every FOCUSS iteration. λ_c is then compared with the limits. If $\lambda_c < \lambda_{min}$ then λ_{min} is used. If $\lambda_c > \lambda_{max}$ then λ_{max} is used. Otherwise λ_c is used. This ensures that the λ will always be acceptable even if there is no distinct L-corner.

3. EXPERIMENTS AND RESULTS

Numerous experiments are conducted on synthetic data to understand the reliability of the methods proposed above. Experiments are done using an 20×30 matrix, $\mathbf{A}$, with random entries chosen from a normal distribution with mean zero and variance one. The columns in $\mathbf{A}$ are normalized. The noise free data vector $\mathbf{b}'$ is obtained as a linear combination of r randomly picked vectors from $\mathbf{A}$ where the coefficients are Gaussian random variables with zero mean and unit variance. The $\mathbf{b}'$ is then normalized. The noisy data vector, $\mathbf{b}$, is $\mathbf{b}' + \mathbf{n}$ where $\mathbf{n}$ is a noise vector with Gaussian random entries with zero mean and variance depending on the Signal to Noise Ratio (SNR) in the experiment. Mathematically, the syn-

Test	p	C	#	$\#r$	$\bar{\epsilon}_1^2$ 10^{-3}	$\bar{\epsilon}^2$ 10^{-3}	ϵ_1^2 / ϵ^2 %
1 F	0	0.8	9.3	5.86	8.0	4.4	57/24
1 O			9.3	5.58	12.3	7.4	29/62
2 F	0		7	5.47	15.0	13.5	38/30
2 O			7	5.29	17.9	16.2	35/43
1 F	0.5	1.2	6.7	5.44	9.6	9.8	30/25
1 O			6.7	5.37	16.9	14.4	33/38
2 F	0.5		7	5.67	9.8	8.6	35/27
2 O			7	5.57	12.4	10.4	37/45
1 F	0.8	1.5	7.9	5.84	8.0	6.3	45/23
1 O			7.9	5.58	12.5	9.4	41/63
2 F	0.8		7	5.70	12.6	11.4	33/24
2 O			7	5.40	16.3	14.6	26/35

Table 1: Experiments done on test1 and test2 with the same data set. SNR= 20 dB, $r = 7$. F:FOCUSS, O:OMP.

Test	p	C	#	$\#r$	$\bar{\epsilon}_1^2$ 10^{-3}	$\bar{\epsilon}^2$ 10^{-3}	ϵ_1^2 / ϵ^2 %
1 F	0	0.8	7.0	4.21	78.2	49.6	55/22
1 O			7.0	4.10	84.0	42.7	32/65
2 F	0		7	4.26	83.9	47.0	56/32
2 O			7	4.13	88.1	42.8	34/58
1 F	0.5	1	5.3	3.99	84.4	73.9	38/50
1 O			5.3	3.78	90.9	70.0	31/38
2 F	0.5		7	4.46	78.5	44.4	15/26
2 O			7	4.30	82.4	40.8	54/62
1 F	0.8	1	6.4	4.28	79.1	58.7	41/19
1 O			6.4	4.09	84.1	51.3	41/63
2 F	0.8		7	4.43	79.8	45.2	39/29
2 O			7	4.18	82.9	44.9	32/42

Table 2: Experiments done on test1 and test2 with the same data set. SNR= 10 dB, $r = 7$. F:FOCUSS, O:OMP.

thetic data can be described as

$$\mathbf{b} = \mathbf{b}' + \mathbf{n}, \text{ where } \mathbf{A}\mathbf{x} = \mathbf{b}'. \quad (7)$$

Each experiment is done with at least 100 different data vectors.

Several factors have been studied to evaluate the experiments. There are two types of error: $\epsilon_1^2 = \|\mathbf{A}\mathbf{x} - \mathbf{b}'\|^2$ and $\epsilon^2 = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$. If one is trying to find the underlying function of a known sparse process, then the first error measure is the most informative. If one is trying to represent a signal in the best possible manner without knowing the underlying generating function (e.g compression), then the latter will be the most informative.

3.1. Discrepancy principle and Sparsity

The Discrepancy principle (test 1) and Sparsity criteria (test 2) are tested on the same dataset, and are therefore evaluated together.

In the experiment of the discrepancy principle, we assume that we know the variance of the noise. That allows bounding of the norm of the error as a function of the noise variance. Let the variance of $n_i, i = 1, 2 \ldots N$ be σ^2. Then $E\|\mathbf{n}\|^2 = N\sigma^2$, and the error bound is set to $CN\sigma^2$, where C is a user chosen factor. When using this approach, the number, r, of vectors from the $\mathbf{A}$ matrix selected to approximate a data vector $\mathbf{b}$ vary for different data vectors. To facilitate a proper comparison, for every data-vector, $\mathbf{b}_i$, the FOCUSS algorithm is run first, and the number of columns selected, r_i, is found. Then the OMP algorithm is run for the same data vector with the restriction that it has to pick exactly the same number (r_i) of vectors. In the experiment of the sparsity criteria, the number of vectors to be selected to approximate the data is fixed. In this experiment the goal is to find the best possible approximation in terms of minimal Mean Square Error (MSE) using a linear combination of r columns from the $\mathbf{A}$ matrix. Unfortunately it is not trivial to control r when FOCUSS is used as the vector selection algorithm. The way it is done in this experiment is the following: For a data-vector $\mathbf{b}_i$, OMP is run first finding an approximation using r vectors. ϵ^2 is calculated and used as an input for the upper bound when running FOCUSS as in the discrepancy principle. Let r_{FOCUSS} be the number of vectors that FOCUSS uses. If $r_{FOCUSS} > r$, it is pruned down to r by using OMP to select r of the r_{FOCUSS} vectors. If $r_{FOCUSS} < r$, extra vectors are added using OMP until r vectors are selected. This way we always use r vectors in each approximation.

The results of the experiments are summarized in Tables 1, and 2. The various parameters that make up this table are as follows: p is a factor in the FOCUSS algorithm, (4), # means the average number of vectors selected per data vector, $\#r$ means the average number of selected vectors which are identical with vectors used to construct $\mathbf{b}'$, $\bar{\epsilon}_1^2$ and $\bar{\epsilon}^2$ is the mean of the error measures, % ϵ_1^2/ϵ^2 means the percentage of the trials where FOCUSS/OMP performs better in terms of ϵ_1^2/ϵ^2. The reason why % ϵ^2 FOCUSS and % ϵ^2 OMP does not add to 1 is that they perform exactly the same for some of the trials. For the experiments with test 1, C is the user chosen factor in the error bound as explained above.

From Table 1, it can be seen that the mean of both ϵ_1^2 and ϵ^2 is less for FOCUSS then OMP in all the experiments. For low SNR in Table 2 this is still the case for the mean of ϵ_1^2 but no longer for ϵ^2. As is seen from Tables 1 and 2, the results of % ϵ_1^2 and % ϵ^2 seems to be in favor of the OMP in many of the experiments, but still the mean values of ϵ_1^2 and ϵ^2 is in favor of the FOCUSS. This is because when OMP performs better it only performs marginally better, but when FOCUSS performs better it sometimes performs significantly better.

3.2. Modified L-curve

The modified L-curve method requires some knowledge of the noise level, or a target on the approximation SNR. In particular, the largest (ϵ_{max}^2) and small-

Test	p	#	$\#r$	$\bar{\epsilon}_1^2$ 10^{-3}	$\bar{\epsilon}^2$ 10^{-3}	ϵ_1^2/ϵ^2 %
SNR 20 dB:						
FOC	0	7.04	5.35	9.7	10.3	53/42
OMP		7.04	5.05	19.2	17.6	47/51
FOC	0.5	6.86	5.32	10.2	9.4	45/36
OMP		6.86	5.05	20.0	17.8	55/55
FOC	0.8	10.69	5.97	8.3	3.6	74/26
OMP		10.69	5.68	11.7	4.9	26/74
SNR 10 dB:						
FOC	0	4.08	3.46	117.1	115.2	52/39
OMP		4.08	3.06	128.3	118.6	39/48
FOC	0.5	4.34	3.58	99.1	93.8	59/41
OMP		4.34	3.22	108.7	116.8	34/46
FOC	0.8	8.38	4.57	82.4	39.3	76/21
OMP		8.38	4.14	93.9	29.5	24/77

Table 3: Experiments done on the modified L-curve method assuming knowledge of the SNR. $r = 7$

est (ϵ_{min}^2) error in the approximation are required. This is used to find $\lambda_{max}, \lambda_{min}$, as described in section 2.3. In each FOCUSS iteration, $\lambda_{max}, \lambda_{min}$, and λ_c are found.

The noise vector $\mathbf{n}$ has Gaussian random entries with variance $\sigma_n^2 = \frac{\text{SNR}}{dim(\mathbf{n})}$, and the SNR level is 10 or 20 dB. $\|\mathbf{n}\|^2$ has a chi-squared distribution and is used to find the limits. The limits ϵ_{min}^2 and ϵ_{max}^2 are chosen as $P(\|\mathbf{n}\|^2 \geq \epsilon_{min}^2) = P(\|\mathbf{n}\|^2 \leq \epsilon_{max}^2) = T$, where T is a chosen threshold. For these experiments a threshold of .1 was used and this gives $\epsilon_{min}^2 = 0.0062$ and $\epsilon_{max}^2 = 0.0142$ for SNR = 20 dB, and 10 times as much for 10 dB.

If the true SNR of the data is unknown, targets for the SNR can be used to decide the error limits. If the desired SNR is approximately X dB, an upper error limit can be set using $X - \Delta_1$ dB as an SNR target, and a lower limit using $X + \Delta_2$ dB.

$$\epsilon_{upper}^2 = 10^{-(X-\Delta_1)/10}\|\mathbf{b}\|^2 \quad (8)$$

$$\epsilon_{lower}^2 = 10^{-(X+\Delta_2)/10}\|\mathbf{b}\|^2 \quad (9)$$

For every data vector, $\mathbf{b}$, an ϵ_{upper}^2 and ϵ_{lower}^2 is calculated using Equation 8, and 9 before the FOCUSS iterations start. For each data vector in the experiment, FOCUSS runs first and r_i is found, then OMP runs on the same data vector and stops after selecting exactly r_i frame vectors. The errors are then compared.

Results from the modified L-curve method are shown for SNR's of 10 and 20 dB in Table 3.

Table 4 shows experiments where the true SNR for the generated data is 20 dB, but assumed to be unknown. A lower target is set to 15 dB and the higher to 25 dB assuming no knowledge about the noise but requiring the approximation to have an SNR between 15 and 25 dB. The results are in favor of the FOCUSS when compared to OMP. The achieved SNR can be calculated from the mean ϵ^2. For $p = 0$ the SNR_{FOCUSS} is 16.8 dB and for $p = 0.5$ it is 17.6 dB.

Test	p	#	$\#r$	$\bar{\epsilon}_1^2$ 10^{-3}	$\bar{\epsilon}^2$ 10^{-3}	ϵ_1^2 / ϵ^2 %
FOC	0	5.48	4.69	20.4	21.1	44/42
OMP		5.48	4.67	22.4	23.1	50/46
FOC	0.5	5.51	4.93	16.3	17.4	50/38
OMP		5.51	4.71	21.8	21.5	45/41

Table 4: Experiments done on the modified L-curve method. $r = 7$, SNR target between 15 and 25 dB. True SNR for generated data is 20 dB

The results has a lower SNR than the true SNR, but the number of selected vectors is approximately 5.5 when $r = 7$ was used to generate the data.

In summary, the original L-curve scheme exerts no strict controls over the approximation quality, and this often results in the regularization parameter improperly choosing between quality of approximation and sparsity, leading to an unreliable procedure. Our proposed scheme remedies this by the requirement of a target SNR, and procedures for determining the target SNR are presented. The target SNR enables setting limits on the SNR desired of the approximations, and then letting the L-curve algorithm find a good trade off between sparsity and quality of fit within the controlled limits ensures robustness. In the context of compression, the possibility of controlling bounds for the error, while obtaining the minimum bit rate at that error level can be a very desirable property.

4. REFERENCES

[1] S. G. Mallat and Z. Zhang, "Matching Pursuits with Time-Frequency Dictionaries," *IEEE Trans. ASSP*, vol. 41, pp. 3397–3415, Dec. 1993.

[2] B. Jeffs, R. Leahy, and M. Singh, "An evaluation of methods for neuromagnetic image reconstruction," *IEEE Biomed.*, vol. BME-34, pp. 713–723, 1987.

[3] I. Gorodnitsky and B. Rao, "Sparse Signal Reconstructions from Limited Data using FOCUSS: A Re-weighted Minimum Norm Algorithm," *IEEE Trans. on Signal Processing*, vol. 45, pp. 600–616, March 1997.

[4] B. D. Rao and K. Kreutz-Delgado, "An Affine Scaling Methodology for Best Basis Selection," *to appear in the IEEE Trans. on Signal Processing*, Jan. 1999.

[5] K. Engan, S. O. Aase, and J. H. Husøy, "Multi-Frame Compression: Theory and Design," *Signal Processing*, vol. 80, (to be published in issue 10) 2000.

[6] B. D. Rao and K. Kreutz-Delgado, "Basis Slection in the presence of Noise," in *Proc. of the 32st Asilomar Conference on Signals, Systems and Computers*, (Monterey, California), Nov. 1998.

[7] P. C. Hansen and D. P. O'Leary, "The Use of the L-Curve in the Regularization of Discrete Ill-posed Problems," *SIAM J. Sct. Comput.*, vol. 14, pp. 1487–1503, Nov. 1993.

[8] B. D. Rao, "Analysis and Extensions of the FOCUSS Algorithm," in *Proc. of the 30th Asilomar Conference on Signals, Systems and Computers*, vol. 2, (Monterey, California), pp. 1218–1223, Nov. 1996.

[9] G. Davis, *Adaptive Nonlinear Approximations*. PhD thesis, New York University, Sept. 1994.

[10] P. C. Hansen, "Analysis of Discrete Ill-posed Problems by Means of the L-curve," *SIAM Review*, vol. 34, pp. 561–580, Dec. 1992.