APPLICATIONS OF THE ℓ_1 NORM IN SIGNAL PROCESSING

James A. Cadzow
Department of Electrical Engineering
Vanderbilt University
Nashville, Tennessee 37235

Abstract

In various data applications, one is confronted with the task of: (i.) finding a best approximate solution to a system of overdetermined linear equations, or, (ii.) finding a best rank q approximation of a matrix. In solving either problem, a sum of squared errors criterion is almost always invoked when determining the *best* solution. In many applications, however, it is preferable to find a solution in which the sum of equation error magnitudes is minimized (i.e., a minimum ℓ_1 norm). This is true for cases in which the data under analysis contains outliers. Unfortunately, there does not exist a closed form solution for the minimum sum of equation error magnitude criterion. One must therefore resort to algorithmic procedures for iteratively finding such a solution. Efficient algorithmic procedures for solving solution either of the problems posed above based are presented and are illustrated by practical examples.

1 INTRODUCTION

Many problems in digital signal processing can be formulated as that of finding a best approximate solution to a system of over-determined linear equations. This is true in such applications as linear prediction, direction finding, exponential modeling, and, image processing. This generic problem can be formulated as

$$\mathbf{y} = A\mathbf{x} + \boldsymbol{\epsilon} \tag{1}$$

where vector $\mathbf{y} \in R^M$ and matrix $A \in R^{M\times N}$ are given. It is then desired to select the *parameter vector* $\mathbf{x} \in R^N$ so as to make the associated *model error vector* $\epsilon \in R^M$ as close to the zero vector as possible. As a measure of best approximate solution, we shall invoke the ℓ_p norm of a vector and matrix. Specifically, the ℓ_p norm of the $M \times N$ matrix A with elements a_{mn} is defined to be

$$\|A\|_p = \left[\sum_{m=1}^{M}\sum_{n=1}^{N} |a_{mn}|^p\right]^{1/p} \tag{2}$$

This matrix size measure satisfies the axioms of a norm so long as the norm index is constrained by $p \geq 1$. Furthermore, it is applicable to row vectors where $M = 1$ and to column vectors where $N = 1$.

There are a variety of methods for obtaining the best approximate solution of the system of linear equations (1) in the ℓ_p norm sense. There are two approaches which will be considered in this paper. The *first method* involves the traditional approach of selecting the parameter vector $\mathbf{x}$ so as to minimize the ℓ_p norm of the model error vector $\boldsymbol{\epsilon}$, that is

$$\textbf{Criterion 1}: \qquad \min_{\mathbf{x}\in R^N} \|\mathbf{y}-A\mathbf{x}\|_p \tag{3}$$

Let an optimal selection be designated by $\mathbf{x}^o$ which may or may not be unique depending on the rank of matrix A and the choice of the norm index p. The nature of an optimal solution is very much dependent on the value assigned to p and the data analyst must be aware of this factor.

In the *second method* for approximating a solution to a system of over-determined linear equations, a

less direct approach in taken in which a reduced rank $q < M$ approximation of an *augmented matrix* is determined, that is

$$\textbf{Criterion 2}: \quad \min_{\substack{B \in R^{M\times(N+1)} \\ rank(B)=q}} \left\| [-\mathbf{y}\vdots A] - B \right\|_p \quad (4)$$

Once an optimum reduced rank augmented matrix is determined, it is then partitioned as $\mathcal{A}^o = [\mathbf{y}^o\vdots A^o]$. Since the columns of the reduced rank augmented matrix approximation are linearly dependent, it follows that the reduced rank system of linear equations $\mathbf{y}^o = A^o\mathbf{x}$ is guaranteed to have a solution. Specifically, any vector whose first component is one will be a solution, that is

$$\begin{bmatrix} 1 \\ \mathbf{x}^o \end{bmatrix} \in \text{null}(B^o) \quad (5)$$

If the null space of matrix $\mathcal{A}^o$is one and a basis vector for the null space has a nonzero first component, then their will be a unique solution of this form. On the other hand, if this null space has dimension greater than one then there will always exist an infinite number of solutions. The data analyst should perform a preliminary experimentation to determine whether the first or second method is more effective for the problem being considered.

2 MINIMUM ℓ_2 NORM SOLUTION TO A SYSTEM OF LINEAR EQUATIONS

The norm index is typically taken to be the least squared error selection $p = 2$ since such a choice leads to a convenient solution for either of these two minimization problems. This desirable attribute is buttressed by the fact that the ℓ_2 choice often leads to acceptable modeling results. It is for these two reasons that the data analyst instinctively appeals to the ℓ_2 criterion for finding a best approximate solution. It is well known that a selection of the parameter vector to minimize criterion (3) for $p = 2$ is given by

$$\textbf{Solution 1}: \quad \mathbf{x} = \mathbf{A}^\dagger\mathbf{y} \quad (6)$$

where $\mathbf{A}^\dagger \in R^{N\times M}$ designates the Moore-Penrose pseudo inverse of matrix A. When matrix A has full column rank N, the Moore-Penrose pseudo matrix is specified by $\mathbf{A}^\dagger = [A^TA]^{-1}A^T$ and in this case there will be a unique solution.

In solving the second minimization problem (4), the singular value decomposition (SVD) of the augmented matrix is first determined, that is

$$[\mathbf{y}\vdots A] = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^T$$

where r denotes the rank of the augmented matrix. In this SVD decomposition, the positive *singular values* σ_k are ordered in the standard monotonically non-increasing fashion (i.e., $\sigma_k \geq \sigma_{k+1}$), the *left singular vectors* $\{\mathbf{u}_k\}$ are orthonormal vectors (i.e., $\mathbf{u}_k^T\mathbf{u}_m = \delta(k-m)$) in R^M and the *right singular vectors* $\{\mathbf{v}_k\}$ are orthonormal vectors (i.e., $\mathbf{v}_k^T\mathbf{v}_m = \delta(k-m)$) in R^{N+1}. The solution to minimization problem (4) is simply obtained by truncating this SVD decomposition to the q outer products associated with the largest singular values, that is

$$\textbf{Solution 2}: \quad B^o = \sum_{k=1}^{q} \sigma_k \mathbf{u}_k \mathbf{v}_k^T \quad (7)$$

We may then appeal to relationship (5) to obtain an approximate solution.

3 MINIMUM ℓ_1 NORM SOLUTION TO A SYSTEM OF LINEAR EQUATIONS

Although the ℓ_2 based modeling problem has the useful attribute of having a closed form solution, it can lead to undesirable modeling results when the data being analyzed possesses a few data outliers. A data outlier is a data point that does not represent the general trend of the data. Data outliers can arise from bad data point recordings or from environmental noise. Whatever the case, these outliers can have a dramatic negative impact on the resultant optimum parameter vector due to the squared error weighting used in ℓ_2 based modeling . When

it is suspected that the data being analyzed contains data outliers, it is useful to use criteria that are not as suspectable to a few data outliers. The ℓ_1 norm is such a criterion since it only weights the magnitude of the error instead of its square.

To illustrate the effect of data outliers, let us consider the case of linear data which is contaminated by additive Gaussian noise as well as containing two data outliers. It is now desired to uncover the linear trend in the noise contaminated data set $y(\Delta), y(2\Delta), \cdots, y(N\Delta)$ where Δ denotes the interval between data points. Using the linear data model $y_M(n) = n\Delta + b$ where Δ and b designate the *slope* and y-axis intercept of the line, a vector representation of the modeling error $\epsilon(n) = y(n) - y_M(n)$ for $n = 1, 2, \cdots, N$ results in the following overdetermined system of linear equations

$$\begin{bmatrix} y(\Delta) \\ y(2\Delta) \\ \vdots \\ y(N\Delta) \end{bmatrix} = \begin{bmatrix} \Delta & 1 \\ 2\Delta & 1 \\ \vdots & \vdots \\ N\Delta & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} + \begin{bmatrix} \epsilon(1) \\ \epsilon(2) \\ \vdots \\ \epsilon(N) \end{bmatrix}$$

If the optimal slope and y-axis intercept parameters are selected so as to minimize criterion (3) for a minimum sum of error magnitudes (i.e., $p = 1$) or a minimum sum of square errors (i.e., $p = 2$) choice, the resultant optimal ℓ_1 and ℓ_2 line fits are shown in Figure 1. It is apparent that the best ℓ_2 line fit is badly skewed by the two data outliers while the best ℓ_1 line fit is relatively unaffected by data outliers and closely approximates the majority of data points. This behavior is characteristic of the ℓ_1 norm criterion and it constitutes a useful tool when a few data outliers are present in the data under analysis.

With the above thoughts in mind, we shall now provide some well-known theorems characterizing ℓ_1 solutions to problem (3) with $p = 1$. The following well-known fundamental theorem describes the minimum ℓ_1 norm solution to a system of M linear equations in N unknowns.

Theorem 1 *Let there be given the column vector* $\mathbf{y} \in \mathbf{R}^{M\times 1}$ *and matrix* $A \in R^{M\times N}$ *which has full column rank* N*. Furthermore, let any subset of* N *rows of matrix* A *be linearly independent. It then follows that there exists a minimum* ℓ_1 *norm approximate solution* $\mathbf{x}^o$

$$\|\mathbf{y} - A\mathbf{x}^o\|_1 = \min_{\mathbf{x}\in R^N} \|\mathbf{y} - A\mathbf{x}\|_1 \tag{8}$$

such that the associated error vector $\mathbf{e}^o = \mathbf{y} - A\mathbf{x}^o$ *has at least* N *of its elements equal to zero.*

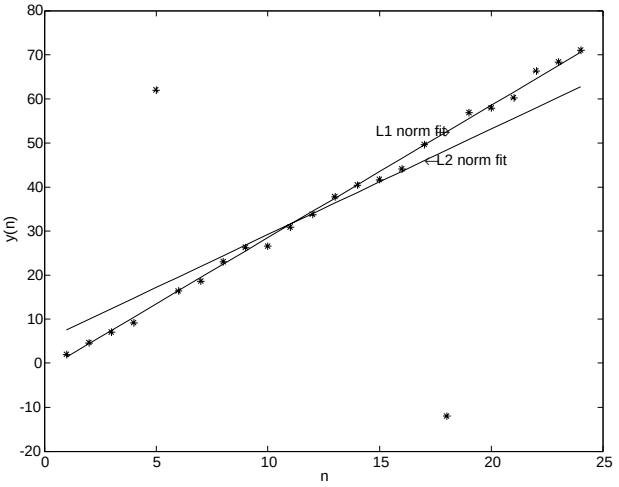

Figure 1: Best ℓ_1 and ℓ_2 line fits to additive noise contaminated linear data.

Conceptually, an optimum ℓ_1 norm solution can be obtained by directly solving each subset of N linear equations from the original system of M linear equations. Each such solution $\mathbf{x}$ is called an *extreme point* and is characterized by the fact at least N components of the associated error vector $\mathbf{e} = \mathbf{y} - A\mathbf{x}$ are zero. Let the set of extreme points be designated by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_s$ where the number of extreme points is bounded above by the *combination* of M things taken N at a time so that $s \leq M!/[N!(M-N)!]$. Theorem 1 indicates that an optimum solution corresponds to any extreme point which renders the measure $\|\mathbf{y} - A\mathbf{x}_k\|_1$ a minimum for $1 \leq k \leq s$. Unfortunately, the need to solve as many as $M!/[N!(M-N)!]$ sets of linear equations makes this direct approach impractical for moderate to large values of M relative to N.

Due to the impracticality of using the direct approach, various algorithms have been developed

which more efficiently use the fact that an extreme point constitutes a solution. The class of *exchange algorithms* are particularly important in this regard. In an exchange algorithm, one equation associated with the prevailing extreme point is exchanged for another equation not associated with the prevailing extreme point to generate a new extreme point. This single equation exchange is strategically made so that the new extreme point has a smaller ℓ_1 norm. This exchange process is continued until an optimum extreme point is eventually obtained. Although more efficient than the direct method, these exchange algorithms are still relatively inefficient since its takes at least one iteration to purge a non-optimum extreme point.

In this paper, a new algorithm is presented whereby the prevailing extreme point is perturbed to an improving point which is typically not an extreme point. By a series of improving perturbations, a new extreme point is then obtained. This process is continued until an optimum extreme point is obtained. This procedure is generally more efficient than an exchange algorithm since it rids itself more rapidly of equations not associated with an optimum extreme point. There exist a number of theorems needed to describe the algorithmic procedure which will be presented at the conference.

4 MINIMUM ℓ_1 NORM LOWER RANK APPROXIMATION OF A MATRIX

The use of the SVD for finding the best rank q approximation of a matrix constitutes a widely used tool for such applications as: (i.) finding a best approximation solution to a system of linear equations, (ii.) decreasing the deleterious effects of additive noise on data, and, (iii.) for image compression. It can happen, however, that the use of a sum of squared error criterion (which is employed in the SVD) for measuring the quality of approximation can lead to yield undesirable results. A sum of magnitude error would be more appropriate in many instances. With this in mind, an effective algorithm for finding the best rank q approximation of the $M \times N$ matrix A in the sense of minimizing the sum of magnitude criterion

$$f(\{\mathbf{u}_k\},\{\mathbf{v}_k\}) = \left\| A - \sum_{k=1}^{q} \mathbf{u}_k \mathbf{v}_k^T \right\|_1 \qquad (9)$$

is described. It is noted that the sum of outer products $\sum_{k=1}^{q} \mathbf{u}_k \mathbf{v}_k^T$ in which the vectors $\mathbf{u}_k \in R^M$ and $\mathbf{v}_k \in R^N$ generates a $M \times N$ matrix with rank less than or equal to r.

Functional (9) is readily shown to be a convex function of the vector set $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q\}$ for a fixed vector set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q\}$ and vice-versa. It is not, however, a convex functional of these two vector sets taken together. Our objective is to select these two vector sets and so as to minimize functional $f(\{\mathbf{u}_k\},\{\mathbf{v}_k\})$. The resultant optimal choice then gives rise to the best rank q matrix approximation as designated by

$$A^{(q)} = \sum_{k=1}^{q} \sigma_k^o \mathbf{u}_k^o \mathbf{v}_k^{oT} \qquad (10)$$

where without loss of generality the optimum vectors are normalized so that $\|\mathbf{u}_k^o\|_1 = \|\mathbf{v}_k^o\|_1 = 1$. An iterative method for finding the best rank q approximation of a matrix is to be presented at the conference. This will include a sequential rank one approximation to initiate the algorithm and the basic steps of the algorithm.

5 ACKNOWLEDGMENT

This work which was in part supported by Amoco Company is heartily acknowledged.

References

[1] Cadzow, James A., *"Application of the ℓ_1 Norm in Signal Processing,"* to be submitted