SOLUTIONS OF UNDERDETERMINED SYSTEMS OF LINEAR EQUATIONS

W. R. Madych*
Department of Mathematics, U-9
University of Connecticut
Storrs, CT 06268

ABSTRACT

We (i) outline a general framework for generating solutions to underdetermined systems of equations, (ii) review properties of several specific methods, including minimal norm and maximum entropy, (iii) introduce specific alternate methods for generating non-negative solutions, (iv) compare, via systematic numerical examples, the solutions generated by these methods with those generated by the maximum entropy and minimum norm methods, and (v) consider the nature of the positivity constraint by studying a transparent example.

* Partially supported by a grant from the Air Force Office of Scientific Research, AFOSR-86-0145.

1980 Math. Subject Classification (1985 Revision). 62G05, 65D99.

1. Introduction

This study is an attempt to obtain some understanding of the nature of certain solutions of underdetermined systems of linear equations with a view to their role in the analysis of various practical inverse problems, specifically those associated with image restoration and computed tomography. For example, one of the questions we are interested in is the following: under what conditions, if any, is the so-called maximum entropy solution better than solutions obtained by other methods. To this end we introduce alternate methods for generating non-negative solutions and discuss the results of several systematic numerical experiments.

More specifically we (i) outline a general framework for generating solutions to underdetermined systems of equations, (ii) review properties of several specific methods, including minimal norm and maximum entropy, (iii) introduce specific alternate methods for generating non-negative solutions, (iv) compare, via systematic numerical examples, the solutions generated by these methods with those generated by the maximum entropy and minimum norm methods, and (v) consider the nature of the positivity constraint by studying a transparent example.

In this paper the basic issue is the non-uniqueness of the solution. We mention but do not specifically address the important issues of noise, ill-conditionedness, and efficiency of the numerical algorithms used to implement the various methods under consideration.

1.1. The basic setup

Consider the system of linear equations

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, i = 1, \ldots, m. \tag{1}$$

Here, we take the field of scalars to be real. Using standard matrix notation (1) may be expressed as a collection of scalar products

$$\langle a_i, x \rangle = b_i, i = 1, \ldots, m, \tag{2}$$

where $a_i = (a_{i1}, \ldots, a_{in})^T$, $x = (x_1, \ldots, x_n)^T$ adn $\langle a_i, x \rangle = a_i^T x$; or even more compactly as

$$Ax = b \tag{3}$$

where $A = (a_{ij})$ is an m by n matrix and $b = (b_1, \ldots, b_m)^T$.

In what follows the matrix A and the data b are assumed to be known. We are interested in the set of solutions to (3), namely the set

$$S_{A,b} = \{x : Ax = b\}.$$

To avoid complications which are not germane to the questions under consideration here, we always assume that $m < n$ and that A is of full rank. Thus $S_{A,b}$ is not empty; indeed, it is an $n - m$ dimensional affine manifold in the real Euclidean space R^n.

1.2. Motivation

In the applications we have in mind, namely image restoration, computed tomography, and related inverse problems, A represents the mathematical model or a discrete analogue of the data acquisition scheme and b is the measured data.

In studying such models many considerations must often be taken into account. For example, in certain models of seismic borehole tomography A is rectangular but not of full rank, see [5]; in this case system (3) is both over and underdetermined. Another complication arises in problems of practical interest by virtue of the fact that there is always some degree of uncertainty in taking physical measurements. Thus in many instances (3) is often replaced with

$$Ax + \varepsilon = b \tag{4}$$

where ε is a random vector which represents noise and has certain statistical properties, see [1], [8], [9], [11]. In view of the fact that there are standard techniques to handle the above complications we feel that taking into account such considerations here will unnecessarily complicate the discussion and cloud the main issue which is the lack of uniqueness.

Returning to the problem modeled by (3), it is clear that the desired quantity is a solution which, unfortunately, is not uniquely determined by the measured data. In fact, the set of feasible solutions $S_{A,b}$ is very large indeed.

Ideally, if one could constrain the feasible set of solutions appropriately, (3) should contain enough information to uniquely determine the x which gave rise to the data. Of course, the best that one can usually expect is to obtain a reasonable estimator.

We outline some general methods for constraining the feasible set of solutions in section two. In the third section we consider some familiar examples, specifically the minimum norm and maximum entropy methods, and introduce some new methods. Finally, in section 4 we consider the consequences of certain constraints, particularly the constraint of componentwise positivity; here we also indicate the results of several numerical experiments.

2. Restricting the feasible set

There are many procedures for restricting the solutions of (3). In this paper we consider two general and often related methodologies. These are described below together with a familiar example. Details of how they are related are contained in subsection 2.3.

2.1. Parametrization

One method is to assume that the solution of (3) is of a certain form. Namely

$$x = F(\xi) \tag{5}$$

where $\xi = (\xi_1, \ldots, \xi_k)$ is contained in a subset of R^k, call it $\mathcal{P}$, and F is a mapping of $\mathcal{P}$ into a subset $\mathcal{M}$ of R^n. Typically the set $\mathcal{P}$ is an open subset of R^k and the mapping F is one to one and continuously differentiable; in this case $\mathcal{M}$ may be viewed as a k dimensional submanifold of R^n. We will always assume that this is the case and, for readily transparent reasons, see (6) below, take $k = m$. Essentially F is a parametrization of the manifold $\mathcal{M}$.

The problem now reduces to finding values of the acceptable parameter ξ such that

$$AF(\xi) = b. \tag{6}$$

If ξ is any solution of (6) then clearly $x = F(\xi)$ is in the intersection of $\mathcal{M}$ and $S_{A,b}$. The main difficulty with this approach in the general case is assuring that the form (5) is such that (6) has a unique solution for every b which may arise in a given application. Furthermore, except for certain examples, it is difficult to determine $\mathcal{M}$, let alone the intersection of $S_{A,b}$ with $\mathcal{M}$.

One classical example where (6) has a unique solution for every b is the case when $\mathcal{P}$ is R^m,

$$x = F(\xi) = \xi_1 a_1 + \ldots + \xi_m a_m = A^T \xi, \tag{7}$$

and $\mathcal{M}$ is the linear subspace generated by $a_1, \ldots, a_m$. In this case the unique solution of (6) is given by

$$\xi = (AA^T)^{-1} b \tag{8}$$

and the corresponding solution x of (3) is given by

$$x = A^T (AA^T)^{-1} b. \tag{9}$$

(We remind the reader that A is assumed to have rank m which implies that AA^T is invertible.) We will return to this important example later.

2.2. Optimization

The other general approach is to find the minimum (or maximum) of a scalar valued function $f(x)$ defined on a subset $\mathcal{K}$ of R^n subject to the constraints imposed by (2). In other words, find x which satisfies

$$f(x) = \min\{f(y) : y \in \mathcal{K} \cap S_{A,b}\}. \tag{10}$$

Of course f should be chosen so that the set $\mathcal{K} \cap S_{A,b}$ is not empty and f has a unique minimum on this set. Fortunately by choosing f with certain readily verifiable properties, for instance, convexity, it is not difficult to guarantee that the desired conditions are satisfied.

For example if f is a positive definite quadratic form on $\mathcal{K} = R^n$ then it is a classical fact that (10) has a unique solution, see [10]. In the special case

$$f(x) = \langle x, x \rangle$$

the solution is known as the *minimum norm solution* of (3) and is given by (9).

2.3. A connection

As mentioned earlier, often the methods of parametrization and optimization are related. To see this, assume that $f(x)$ is continuously differentiable and formally apply the method of Lagrange multipliers to solve problem (10). In particular, set $\xi = (\xi_1, \ldots, \xi_m)^T$ where the ξ_i are the "lagrange multipliers" and write

$$h(\xi, x) = f(x) - \sum_{i=1}^{m} \xi_i (\langle a_i, x \rangle - b_i). \tag{11}$$

Taking the gradient of h and setting the result to 0 gives

$$\langle a_i, x \rangle = b_i, i = 1, \ldots, m,$$

and

$$\frac{\partial f(x)}{\partial x_j} - \sum_{i=1}^{n} \xi_i a_{ij} = 0, j = 1, \ldots, n.$$

The first set of equations is just (2). The last set of equations can be written more compactly and transparently as

$$\nabla f(x) = A^T \xi \tag{12}$$

where ∇f denotes the gradient of f. If ∇f is invertible the optimal solution may be expressed as

$$x = G(A^T \xi) \tag{13}$$

where G is the inverse of ∇f. Finally, equations (2) and (13) imply that under appropriate conditions the solution to problem (10) is given by x of the form (13) where ξ is a solution of

$$AG(A^T \xi) = b. \tag{14}$$

The relationship mentioned above should now be clear. Roughly speaking, if $F(\xi) = G(A^T \xi)$, where G is the inverse of ∇f then the two methods should give rise to the same solution. This observation is useful when trying to determine whether (5) has a unique solution and other questions related to F.

In the case where f is of the form

$$f(x) = \sum_{j=1}^{n} f_j(x_j) \tag{15}$$

equation (13) is particularly simple. Namely, it can be expressed as

$$x_j = g_j((A^T \xi)_j), j = 1, \ldots, n, \tag{16}$$

where $(A^T \xi)_j$ denotes the j-th component of $A^T \xi$ and g_j is the inverse of the (univariate) derivative of f_j. Essentially all the examples below are of this form.

3. Examples

3.1. Minimum norm and generalizations

As indicated earlier the minimum norm solution of (3) is the solution to problem (10) in the case

$$f(x) = \langle x, x \rangle \tag{17}$$

perhaps a more accurate description would be the minimum quadratic norm solution.

Standard variants of (17) are more general quadratics which include f's of the form

$$f(x) = q(C(x - y))$$

where $q(x) = \langle x, x \rangle$, C is an $n \times n$ matrix, and y is a constant vector. The parameters in C and y are usually chosen to influence the behavior of the solution. Properties of such solutions are well known and well documented, for example see [2], [3], [10], and the references cited there.

Perhaps the most important feature of this method is the fact that the relationship between the data and the resulting estimator is linear. Besides being a convenient theoretical tool, this property allows for efficient computational algorithms in many applications.

3.2. Maximum entropy

The maximum entropy solution of (1) is an estimator whose components are of the form

$$x_j = p_j \exp\left(\sum_{i=1}^{n} \xi_i a_{ij}\right), j = 1, \ldots, n, \tag{18}$$

where $p = (p_1, \ldots, p_n)^T$ is a constant and the ξ_i's are parameters chosen so that $x = (x_1, \ldots, x_n)^T$ satisfies (1). Using the notation and terminology of subsection 2.1 this form can be expressed more compactly as

$$x = F(\xi) = P \exp(A^T \xi), \tag{19}$$

where P si the constant diagonal matrix with diagonal p and the exponential is interpreted componentwise, namely, if $y = (y_1, \ldots, y_n)^T$ then $\exp(y) = (\exp(y_1), \ldots, \exp(y_n))^T$.

Observe that in this case the manifold $\mathcal{M}$ is contained in the positive cone

$$R_n^+ = \{x : x_j > 0, j = 1, \ldots, n\}.$$

Thus it should be clear that any estimator of this form will have non-negative components.

The parameter p is chosen to influence the behavior of the solution. In the discussion below, unless indicated otherwise, we always assume the components of p to be one.

Observe that the manifold $\mathcal{M}$ can be described so that in low dimensional examples it is relatively easy to visualize. For example, in the case $m = 2$, $n = 3$, if the last column of A is a linear combination of the first and second with coefficients α_1 and α_2 respectively, then $\mathcal{M}$ may be described by

$$\mathcal{M} = \{x : x_1 > 0, x_2 > 0, x_3 = x_1^{\alpha_1} x_2^{\alpha_2}\}.$$

In the general case, if A is such that the first m columns are linearly independent, and the components of x can always be permuted so that the resulting system of equations has this property, then $\mathcal{M}$ is the intersection of the manifolds $\mathcal{M}_i, i = 1, \ldots, n - m$, where

$$\mathcal{M}_i = \{x : x_1 > 0, \ldots, x_m > 0, x_{m+i} = x_1^{\alpha_{i1}} \cdots x_m^{\alpha_{im}}\}.$$

and the α_{ij}'s are appropriate constants.

It is not difficult to see that the resulting estimator may also be viewed as the solution of the minimization problem (10) with

$$f(x) = \sum_{j=1}^{n} x_j \log\left(\frac{x_j}{e p_j}\right), \tag{20}$$

where log is the natural logarithm with base e and the constants p_j are those in (18). As suggested in subsection 2.3, an immediate consequence of this formulation is the fact that if the solution hyperplane $S_{A,b}$ intersects the positive cone R_+^n then $S_{A,b}$ intersects the manifold parametrized by (19) at exactly one point. In other words, if $R_+^n \cap S_{A,b}$ is not empty then (3) has a unique solution of the form (19).

Note that the particular normalization of f in (20) results in $x = p$ as the optimal solution in the case of no constraints.

It should be mentioned that the maximum entropy solution is often viewed as that estimator which maximizes the negative of an expression similar to (20). Thus the term maximum. The rationale behind the term entropy is discussed in [4] and [6].

In view of the fact that much work has centered around this method surprisingly little is known concerning the theoretical properties of the resulting estimators beyond the immediate consequences of the definitions.

One interesting fact concerning such solutions to a very special class of linear systems has been given in [4] and [7]. These systems can be described as follows: Suppose $n = lk$ and write the variable $x = (x_1, \ldots, x_n)$ in a rectangular array as shown.

$$\begin{array}{ccc} x_1 & \cdots & x_k \\ \vdots & & \vdots \\ x_{lk-k+1} & \cdots & x_n \end{array} \tag{21}$$

The system of equations then is simply the collection of row sums and column sums of (21), namely

$$\begin{aligned} \sum_{j=1}^{k} x_{(i-1)k+j} &= r_i, i = 1, \ldots, l, \\ \sum_{i=1}^{l} x_{(i-1)k+j} &= c_j, j = 1, \ldots, k, \end{aligned} \tag{22}$$

where each row and column sum is positive and

$$r_1 + \ldots + r_l = 1$$

$$c_1 + \ldots + c_k = 1.$$

For this system the maximum entropy solution is given by

$$x_{(i-1)k+j} = r_i c_j. \tag{23}$$

In other words, the value of each component is the product of the row and column sums which contain it.

The setup involving (21), (22), and (23) has the following probabilistic interpretation. If the x_j's represent the probabilities of certain basic events $\omega_j, j = 1, \ldots, n$, then the sums in (22) represent the probabilities of the unions,

$$\rho_i = \cup_{j=1}^{k} \omega_{(i-1)k+j}, i = 1, \ldots, l \text{ and } \gamma_j = \cup_{i=1}^{l} \omega_{(i-1)k+j}, j = 1, \ldots, k.$$

Formula (23) expresses the fact that ρ_i and γ_j are mutually independent events in the probabilistic sense. This interpretation is valid only for systems of the type described by (22); presently there are no analogous results for more general systems.

3.3. Methods for generating bounded solutions

In this subsection we introduce two alternate methods for generating solutions which are bounded componentwise based on the generalities in the second section.

Observe that it suffices to restrict our attention to those methods which generate estimators which are in the positive cone R^n_+ since the change of variables $x \longrightarrow x - y$ will easily transform such a method to one which generates estimators which are bounded componentwise from below by y. A similar remark holds concerning boundedness from above.

If $\xi = (\xi_1, \ldots, \xi_m)^T, -\infty < \xi_i < \infty, i = 1, \ldots, m$, consider the parametrization given by

$$x_j = \frac{1}{2}\{(A^T\xi)_j + \sqrt{(A^T\xi)_j^2 + 4}\}, j = 1, \ldots, n, \tag{24}$$

where $(A^T\xi)_j$ denotes the j-th component of $A^T\xi$. This parametrization can be expressed more compactly by the formula

$$x = \Phi(A^T\xi), \tag{25}$$

where if $y = (y_1, \ldots, y_n)^T$ then the j-th component of Φ is given by

$$\Phi_j(y) = \frac{1}{2}\{y_j + \sqrt{y_j^2 + 4}\}, j = 1, \ldots, n.$$

Observe that if $\mathcal{M}_\Phi$ denotes the manifold parametrized by (24) then $\mathcal{M}_\Phi$ is contained in the positive cone R^n_+. Thus any solution of (3) which enjoys the representation (24) must have non-negative components.

Proposition 1. *If the intersection of $S_{A,b}$ and R^n_+ is not empty then*

$$A\Phi(A^T\xi) = b \tag{26}$$

has a unique solution $\hat{\xi}$. In other words, $\mathcal{M}_\Phi \cap S_{A,b}$ contains exactly one element $\hat{x} = \Phi(A^T\hat{\xi})$.

Proof. Consider the scalar function f defined on R^n_+ by the formula

$$f(x) = \sum_{j=1}^{n} \left(\frac{x_j^2}{2} - \log x_j \right). \tag{27}$$

in R^n_+. Using the chain of reasoning outlined in subsection 2.3 it is apparent that such an $\hat{x}$ satisfies

$$\nabla f(\hat{x}) = A^T\hat{\xi}$$

for a unique $\hat{\xi}$. Since Φ is the inverse of ∇f we may write

$$\hat{x} = \Phi(A^T\hat{\xi})$$

and the desired result follows. □

For another example consider the relation

$$s = t - \frac{1}{t^3} \tag{28}$$

for positive numbers t. Since the right hand side of (28) is an increasing function of t it is evident that (28) defines a function $\psi(s)$ mapping $0 < s < \infty$ onto $-\infty < t < \infty$. Note that $\psi(s)$ is a root of the polynomial $t^4 - st^3 - 1$. Now, using roughly the same notation as in (25), we define another parametrization by the formula

$$x = \Psi(A^T\xi), \tag{29}$$

where the j-th component of Ψ is given by

$$\Psi_j(y) = \psi(y_j), j = 1, \ldots, n,$$

and ψ is the function defined earlier in this paragraph.

It is not difficult to see that everything that was said concerning Φ also holds for Ψ. Indeed, we may state the following.

Proposition 2. *Proposition 1 remains true if Φ is replaced by Φ.*

Proof. Recall the proof of Proposition 1. if we simply replace (27) by

$$f(x) = \sum_{j=1}^{n} \frac{1}{2}\left(x_j^2 + \frac{1}{x_j^2}\right) \tag{30}$$

the rest of the proof of this proposition is the same as that of Proposition 1. □

A direct consequence of the above propositions are two methods for generating estimators for (3). For future reference we will refer to them as method one and method two.

4. Numerical experiments and comments

Recall that the estimators generated by the method of minimum norm are linear functions of the data. Furthermore there are a host of efficient algorithms for computing them, see [2] and [10]. On the other hand, the estimators generated by the other methods mentioned above depend non-linearly on the data and, as is readily evident, are considerably more difficult to compute. Thus understanding the nature of these estimators and their relative merits is of some practical significance.

4.1. Description of numerical experiments

To obtain a sense of the nature of the estimators generated by the methods outlined in Section 3 we performed numerical experiments on relatively small linear systems. The systems considered were of the form

$$\frac{1}{k}\sum_{j=0}^{k-1} x_{i+j} = b_i, i = 1, \ldots, m, \tag{31}$$

where $k < n, m = n-k+1$, and n is the number of variables x_j. System (31) is a typical example of a one dimensional blurring model. The sizes considered for our experiments ranged from $m = 15$, $n = 20$ to $m = 80$, $n = 100$ with the ratio m/n varying between 0.5 and 0.9. The experiments were performed as follows:

A non-negative pseudo-random vector $x = (x_1, \ldots, x_n)^T$ was generated via a canned subroutine and the data $b = Ax$ was computed. Then each of the methods outlined in Section 3 was used to generate an estimator. We refer to those methods as minimum norm, maximum entropy, method one, and method two or, more briefly, MN, ME, M1, and M2 respectively. Formula (9) was used to compute the MN estimator. The other estimators were computed by using Newton's method to solve (6) for $\hat{\xi}$ and then using (5) to evaluate the corresponding estimator $\hat{x}$. In the case of maximum entropy the values $p_j = 1, j = 1, \ldots, n$, were used. The resulting estimators were plotted together with the true phantom x. In each case the error

$$\|x - \hat{x}\| = \left\{ \sum_{j=1}^{n} |x_j - \hat{x}_j|^2 \right\}^{1/2}$$

was computed.

The results of these experiments can be loosely described as follows:

When the lower bound, zero in this case, was not a tight one for the phantom then all the methods generated estimates which were roughly equivalent. Namely, they differed from one another but the differences were judged not significant. The computed error varied but was roughly the same order of magnitude for all the methods. For example, see Figure 1; here $m = 19$, $n = 24$ and the components of the phantom x are uniformly distributed between 0 and 1.

On the other hand, when the lower bound for the phantom was reasonably tight then the methods which enforced the lower bound generated considerably better estimators. Indeed, the computed errors for estimators generated by ME, M1 and M2 were significantly smaller than the error for the estimator generated by MN. The estimators generated by ME, M1, and M2 differed but the difference was judged not significant; the same was true of the corresponding computed error. For example, see Figure 2; here again $m = 19, 4n = 24$ but the components of the phantom x are distributed according to the 1/4-th power of the uniform distribution between 0 and 1.

Similar experiments were performed on systems of the form

$$\sum_{j=1}^{n} x_j \cos \frac{2\pi ij}{n}, i = 0, \ldots, k,$$

$$\sum_{j=1}^{n} x_j \sin \frac{2\pi ij}{n}, i = 1, \ldots, k,$$

where $m = 2k + 1$ and m significantly less than n. Here the results were roughly the same as those reported above, although in the cases when the lower bound was tight on the phantom the differences seemed less dramatic.

4.2. Comments

To obtain some perspective on the observations recorded in the previous subsection consider the system given by

$$x_1 + x_2 = \varepsilon \text{ and } x_2 + x_3 = 1 \tag{32}$$

where ε is a small positive number. In this case $S_{A,b}$ is simply the line in R^3 parametrized by

$$x = (\varepsilon - t, t, 1 - t), -\infty < t < \infty.$$

Suppose that, in addition to (32) we know that the phantom which gave rise to the data had non-negative components. Namely

$$x_1 \geq 0, x_2 \geq 0, x_3 \geq 0. \tag{33}$$

It is then clear that with this additional information the feasible set of solutions is that part of $S_{A,b}$ contained in a ball of radius $\sqrt{3}\varepsilon/2$ centered at $(\varepsilon/2, \varepsilon/2, 1 - (\varepsilon/2))$. Thus any non-negative solution of (32) will be within $\sqrt{3}\varepsilon$ of the desired phantom. Indeed, if $\varepsilon = 0$, then (32) together with (33) imply a unique solution.

The principle illustrated by the above simple example is no doubt valid for much larger and more complicated systems of equations. The extent to which this principle is valid depends on the matrix A and the phantom x and should be possible to characterize quantitatively in terms of these parameters.

References

[1] Geman, S., & McClure, D. E. (1985). Bayesian image analysis: An application to single photon emission tomography. *Proceedings of American Statistical Association, Statistical Computing Section*, 12–18.

[2] Golub, G., & Van Loan, C. (1983). *Matrix Computations.* John Hopkins, Baltimore.

[3] Golomb, M., & Weinberger, H. F. (1959). Optimal Approximation and Error Bounds. Langer, R. E. (1959, ed.), *On Numerical Approximation*, Madison, 117–190.

[4] Gull, S. F., & Skilling, J. (1984). Maximum entropy method in image processing. *IEE Proceedings* **131**, 646–659.

[5] Ivanson, S. (1986). Seismic Borehole Tomography–Theory and Computational Methods. *Proceedings of the IEEE* **74**, 328–338.

[6] Jaynes, E. T. (1983). *Papers on probability statistics and statistical physics*, R. Rosencrantz, ed., Reidel, Dordrecht.

[7] Livesey, A. K., & Skilling, J. (1985). Maximum Entropy Theory. *Acta Cryst.* A **41**, 113-122.

[8] Madych, W. R., & Nelson, S. A. (1983). Polynomial based algorithms for computed tomography. *SIAM Journal of Applied Mathematics* **43**, 157–185.

[9] Shepp, L. A., & Vardi, Y. (1982). Maximum likelihood reconstruction in positron emission tomography. *IEEE Transactions on Medical Imaging* **1**, 113–122.

[10] Strang, G. (1986). *Introduction to Applied Mathematics*, Wellesley–Cambridge, Wellesley.

[11] Titterington, D. M. (1987). Regularisation procedures in signal processing and statistics. Durrani, T. S. (1987, ed.), *Mathematics in Signal Processing*, 213–223, Clarendon, Oxford.