SIAM J. NUMER. ANAL.
Vol. 16, No. 6, December 1979

BOUNDS ON SOLUTIONS OF LINEAR SYSTEMS WITH INACCURATE DATA*

J. E. COPE† AND B. W. RUST†

Abstract. Oettli, Prager, and Wilkinson (1965), (1964), (1965) have dealt with the problem of finding the solution set of a system of m equations in n unknowns,

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

where $\mathbf{A}$ and $\mathbf{b}$ are known only to some limited tolerance, $\mathbf{A}^0-\Delta\mathbf{A}\leqq\mathbf{A}\leqq\mathbf{A}^0+\Delta\mathbf{A}$, $\mathbf{b}^0-\Delta\mathbf{b}\leqq\mathbf{b}\leqq\mathbf{b}^0+\Delta\mathbf{b}$ with $\Delta\mathbf{A}$ and $\Delta\mathbf{b}$ being arrays consisting of positive elements. Given an n-vector $\mathbf{x}$, necessary and sufficient conditions are established in [3] for $\mathbf{x}$ to be an exact solution for some system $\bar{\mathbf{A}}\mathbf{x}=\bar{\mathbf{b}}$, where $\mathbf{A}^0-\Delta\mathbf{A}\leqq\bar{\mathbf{A}}\leqq\mathbf{A}^0+\Delta\mathbf{A}$ and $\mathbf{b}^0-\Delta\mathbf{b}\leqq\bar{\mathbf{b}}\leqq\mathbf{b}^0+\Delta\mathbf{b}$. These results are extended and the emphasis changed. If it is known from some *a priori* consideration the orthant in which the true solution vector $\mathbf{x}$ lies, and if $\mathbf{A}$ and $\mathbf{b}$ are as above, it is possible to compute bounds for $\mathbf{x}$ by linear programming. The tableau for the problem is developed. The results are extended to finding confidence intervals for $\mathbf{x}$ in the case where $\mathbf{A}$ and $\mathbf{b}$ are random samples from distributions with known variances. A discussion of the application of the technique to solving first kind Fredholm equations is also given. The technique is applied to three example problems.

1. Introduction and development of method. In this paper we show how to compute bounds on the solution set of a system of m equations with n unknowns,

$$\mathbf{A}\mathbf{x}=\mathbf{b}. \tag{1.1}$$

The elements a_{ij} of the coefficient matrix $\mathbf{A}$, and the elements b_i of the right-hand side $\mathbf{b}$, are known only to limited accuracy, and may take any values within the intervals

$$\begin{aligned} a^0_{ij}-\Delta a_{ij}\leqq a_{ij}\leqq a^0_{ij}+\Delta a_{ij},\\ b^0_i-\Delta b_i\leqq b_i\leqq b^0_i+\Delta b_i\end{aligned} \tag{1.2}$$

where Δa_{ij} and Δb_i are all nonnegative tolerances. In order to simplify later notation, we rewrite (1.2) as

$$\begin{aligned} \mathbf{A}^0-\Delta\mathbf{A}\leqq\mathbf{A}\leqq\mathbf{A}^0+\Delta\mathbf{A},\\ \mathbf{b}^0-\Delta\mathbf{b}\leqq\mathbf{b}\leqq\mathbf{b}^0+\Delta\mathbf{b}\end{aligned} \tag{1.2$'$}$$

or

$$\begin{aligned} \mathbf{A}\in\mathcal{A}=\{\mathbf{A}|\mathbf{A}^0-\Delta\mathbf{A}\leqq\mathbf{A}\leqq\mathbf{A}^0+\Delta\mathbf{A}\}\\ \mathbf{b}\in\mathcal{B}=\{\mathbf{b}|\mathbf{b}^0-\Delta\mathbf{b}\leqq\mathbf{b}\leqq\mathbf{b}^0+\Delta\mathbf{b}\}.\end{aligned} \tag{1.2$''$}$$

An n-vector $\mathbf{x}$ is said to be a solution of the system (1.1) if there exist δa_{ij} and δb_i with $|\delta a_{ij}|\leqq\Delta a_{ij}$, $|\delta b_i|\leqq\Delta b_i$, $i=1,\cdots,m$, $j=1,\cdots,n$, such that $\mathbf{x}$ is an exact solution of the system $(\mathbf{A}^0+\delta\mathbf{A})\mathbf{x}=\mathbf{b}^0+\delta\mathbf{b}$. It is shown in [3] that the above condition holds if and only if

$$\sum_{j=1}^{n}\Delta a_{ij}|x_j|+\Delta b_i\geqq\Big|\sum_{j=1}^{n}a^0_{ij}x_j-b^0_i\Big|,\qquad i=1,\cdots,n. \tag{1.3}$$

In [3], Oettli shows that if the Δa_{ij} and Δb_i are small enough so that all solutions lie in the same orthant, then bounds on the possible solutions can be found. We extend the results of [3], [4], [5] and change the approach to the problem. Instead of requiring that $\Delta\mathbf{A}$ and

* Received by the editors August 22, 1977, and in final revised form September 6, 1978.

† Union Carbide Corporation, Nuclear Division, Oak Ridge, Tennessee 37830.

$\Delta\mathbf{b}$ remain small, we allow the solution set to extend over many orthants, or even be unbounded, and look for bounds on all of the solutions which lie within a given orthant. The reasons for this point of view are that there is no a priori way to tell how small $\Delta\mathbf{A}$ and $\Delta\mathbf{b}$ must be to keep all solutions in the same orthant, and that often nothing can be done about uncertainty in $\mathbf{A}$ and $\mathbf{b}$. Hence, instead of keeping $\Delta\mathbf{A}$ and $\Delta\mathbf{b}$ small enough to keep the solutions from going outside a given orthant, we do not consider any solutions lying outside that orthant. The problem becomes the following: given $\mathbf{A}$ and $\mathbf{b}$ as above, find in the given orthant, values $x_j^{lo} \leqq x_j^{hi}$, $j = 1, \cdots, n$, such that for all solutions $\mathbf{x}$ in that orthant to the systems defined by (1.1) and (1.2), $x_j \in [x_j^{lo}, x_j^{hi}]$. Although we do not know $\mathbf{A}$ and $\mathbf{b}$ exactly, we assume that an exact $\bar{\mathbf{A}}$ and $\bar{\mathbf{b}}$ exist, and that $\mathbf{A}^0 - \Delta\mathbf{A} \leqq \bar{\mathbf{A}} \leqq \mathbf{A}^0 + \Delta\mathbf{A}$, $\mathbf{b}^0 - \Delta\mathbf{b} \leqq \bar{\mathbf{b}} \leqq \mathbf{b}^0 + \Delta\mathbf{b}$. We assume that the system $\bar{\mathbf{A}}\mathbf{x} = \bar{\mathbf{b}}$ has at least one solution $\mathbf{x}$, and that we know in which orthant $\mathbf{x}$ lies. We then find bounds on all solutions within that orthant.

Knowing the orthant of the solution vector is not so artificial as it may seem, since in many applied problems, a priori physical constraints dictate the signs of the components x_j. The system $\mathbf{A}\mathbf{x} = \mathbf{b}$ may have solutions in other orthants, but those are not considered in computing bounds for $\mathbf{x}$. We note that if $\Delta\mathbf{A}$ is large enough and if there is no orthant constraint, the solution set may become unbounded. Thus the assumption of an orthant constraint replaces the requirement that $\Delta\mathbf{A}$ and $\Delta\mathbf{b}$ be small. Of course it is possible to have problems in which an orthant constraint is not powerful enough to bound the solution set.

Using (1.3) we set up the tableaux for the linear programming problems which must be solved to compute the bounds for the solution. We can write

$$|x| = q, \tag{1.4}$$

where q is an n-vector, $q_j = |x_j|$.

Rewrite (1.3) as two inequalities,

$$\sum_j \Delta a_{ij}|x_j| + \Delta b_i \geqq \sum_j a_{ij}^0 x_j - b_i^0, \tag{1.5a}$$

$$\sum_j \Delta a_{ij}|x_j| + \Delta b_i \geqq -\sum_j a_{ij}^0 x_j + b_i^0. \tag{1.5b}$$

Let sgn $x_j = 1$ if $x_j \geqq 0$, and sgn $x_j = -1$ if $x_j < 0$. Since the orthant of $\mathbf{x}$ is known, sgn x_j is known, and $x_j = \operatorname{sgn} x_j |x_j|$, so (1.5a) and (1.5b) may be rewritten

$$\sum_j \Delta a_{ij}|x_j| + \Delta b_i \geqq \sum_j a_{ij}^0 \operatorname{sgn} x_j |x_j| - b_i^0, \tag{1.6a}$$

$$\sum_j \Delta a_{ij}|x_j| + \Delta b_i \geqq -\sum_j a_{ij}^0 \operatorname{sgn} x_j |x_j| + b_i^0. \tag{1.6b}$$

Since $\mathbf{q} = |\mathbf{x}|$, we have, in matrix notation, where $\mathbf{Sg} = \operatorname{diag}(\operatorname{sgn} x_1, \cdots, \operatorname{sgn} x_n)$,

$$\begin{pmatrix} \mathbf{A}^0\mathbf{S}_g - \Delta\mathbf{A} \\ -\mathbf{A}^0\mathbf{S}_g - \Delta\mathbf{A} \end{pmatrix} \mathbf{q} \leqq \begin{pmatrix} \mathbf{b}^0 + \Delta\mathbf{b} \\ -\mathbf{b}^0 + \Delta\mathbf{b} \end{pmatrix}. \tag{1.7}$$

We wish to find the minimum and maximum possible values of x_j so that $\mathbf{x}$ will satisfy (1.3) and the orthant constraint. This can be done by solving a series of $2n$ linear programming problems, each having the constraint region defined by (1.7) and a nonnegativity constraint on the vector $\mathbf{q}$. To find the upper bound for x_j choose the cost vector to be $\mathbf{c}_j = (0, \cdots, 0, \operatorname{sgn} x_j, 0, \cdots, 0)^T$ and maximize the quantity $\mathbf{c}_j^T\mathbf{q}$. Similarly to find the lower bound for x_j maximize the quantity $-\mathbf{c}_j^T\mathbf{q}$. The bounds on $\mathbf{x}$ are then found by solving the $2n$ linear programming problems:

$$(1.7)' \qquad \begin{matrix} x_j^{hi} \\ x_j^{lo} \end{matrix} = \max_{q} \left\{ \begin{matrix} \mathbf{c}_j^T\mathbf{q} \\ -\mathbf{c}_j^T\mathbf{q} \end{matrix} \middle| \begin{pmatrix} \mathbf{A}^0\mathbf{S}_g - \Delta\mathbf{A} \\ -\mathbf{A}^0\mathbf{S}_g - \Delta\mathbf{A} \end{pmatrix} \mathbf{q} \leqq \begin{pmatrix} \mathbf{b}^0 + \Delta\mathbf{b} \\ -\mathbf{b}^0 + \Delta\mathbf{b} \end{pmatrix}, \mathbf{q} \geqq 0 \right\}.$$

The solutions yield lower bounds x_j^{lo} and upper bounds x_j^{hi} for the components x_j of a solution vector. Hence, the vector interval $I = [\mathbf{x}^{lo}, \mathbf{x}^{hi}] \subseteq R^n$ contains the solution set to the system $\mathbf{Ax} = \mathbf{b}$ (except, of course, for solutions which do not lie in the proper orthant). Note, however, that our results give only maxima and minima on *each component* of $\mathbf{x}$; they do not imply that every vector $x \in I$ is a solution. A priori information about $\mathbf{x}$, in addition to the orthant constraint, can sometimes be incorporated into the tableau. For example, suppose it is known that $x_1 \geqq x_2 \geqq \cdots \geqq x_n \geqq 0$, i.e. $\mathbf{x}$ is monotonically nonincreasing and nonnegative. Then we can write $\mathbf{x} = \mathbf{Rq}$, where $q_j \geqq 0$, $j = 1, \cdots,$ and

$$R = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix},$$

i.e. $\mathbf{R}$ is the upper triangular matrix with $R_{ij} = 1$ for $i \leqq j$. Similarly, if $0 \leqq x_1 \leqq \cdots \leqq x_n$ then $\mathbf{x} = \mathbf{Rq}$, where $\mathbf{R}$ is the lower triangular matrix with all ones. We then have the system $\mathbf{ARq} = \mathbf{b}$, and we wish to find $\begin{Bmatrix} \min \\ \max \end{Bmatrix} \mathbf{c}^T(\mathbf{Rq})$, i.e., our new cost vector is $\mathbf{c}^T\mathbf{R}$. The constraints are then

$$\begin{pmatrix} \mathbf{A}^0\mathbf{S}g\mathbf{R} - \Delta\mathbf{AR} \\ -\mathbf{A}^0\mathbf{S}_g\mathbf{R} - \Delta\mathbf{AR} \end{pmatrix} \mathbf{q} \leqq \begin{pmatrix} \mathbf{b}^0 + \Delta\mathbf{b} \\ -\mathbf{b}^0 + \Delta\mathbf{b} \end{pmatrix}, \qquad \mathbf{q} \geqq \mathbf{0},$$

and we seek $\max\{\mathbf{c}^T\mathbf{Rq}\}$ and $\max\{-\mathbf{c}^T\mathbf{Rq}\}$. If desired, further information on bounds for linear combinations of the x_j may be obtained by using suitable choices of $\mathbf{R}$; for example, $(x_1 + x_2)^{hi}$ may be computed by taking

$$R = \begin{pmatrix} 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix},$$

and using the cost vectors $\mathbf{c}_1^T\mathbf{R}$. Similarly, $(x_1 + x_2)^{lo}$ may be computed by taking the vectors $-\mathbf{c}_1^T\mathbf{R}$.

Up to now, we have assumed that although there was uncertainty about the exact values of $\mathbf{A}$ and $\mathbf{b}$, we knew for sure that they lay within certain bounds (1.2), i.e.,

$$\Pr\{\bar{\mathbf{A}} \in \mathscr{A}\} = 1 \quad \text{and} \quad \Pr\{\bar{\mathbf{b}} \in \mathscr{B}\} = 1$$

Now suppose that we know only confidence intervals for $\bar{\mathbf{A}}$ and $\bar{\mathbf{b}}$, that is,

$$(1.8) \qquad \Pr\{\bar{\mathbf{A}} \in \mathscr{A}\} = \alpha \leqq 1 \quad \text{and} \quad \Pr\{\bar{\mathbf{b}} \in \mathscr{B}\} = \alpha' \leqq 1.$$

We wish to establish probabilistic bounds for $\mathbf{x}$. Suppose (1.8) holds and compute $\mathbf{x}^{lo}$

and $\mathbf{x}^{hi}$ as above. Now suppose that $\bar{\mathbf{x}}$ is a solution of the system $\mathbf{Ax}=\mathbf{b}$, lying in the proper orthant. If $\bar{\mathbf{A}} \in \mathscr{A}$ and $\bar{\mathbf{b}} \in \mathscr{B}$, then $\bar{\mathbf{x}} \in [\mathbf{x}^{lo}, \mathbf{x}^{hi}]$ by the preceding results. Hence the probability that $\bar{\mathbf{x}}$ lies in the interval $[\mathbf{x}^{lo}, \mathbf{x}^{hi}]$ is at least as large as the combined probability that $\bar{\mathbf{A}} \in \mathscr{A}$ and $\bar{\mathbf{b}} \in \mathscr{B}$. Since $\Pr\{\bar{\mathbf{A}} \in \mathscr{A}\}$ and $\Pr\{\bar{\mathbf{b}} \in \mathscr{B}\}$ are assumed independent,

$$\Pr\{\bar{\mathbf{x}} \in [\mathbf{x}^{lo}, \mathbf{x}^{hi}]\} \geqq \alpha\alpha'. \tag{1.9}$$

In summary, we have shown that if we are given a system $\mathbf{Ax}=\mathbf{b}$ with uncertainties in $\mathbf{A}$ and $\mathbf{b}$, and if we know the orthant of the true solution vector $\bar{\mathbf{x}}$, we can find bounds for the solution vector. Further, if a confidence interval for $\mathbf{A}$ and a confidence interval for $\mathbf{b}$ are known, a corresponding confidence interval can be established for $\mathbf{x}$. Note that the sizes of Δa_{ij} and Δb_i do not affect the problem, except in widening the bounds on $\mathbf{x}$. It is not necessary for Δa_{ij} and Δb_i to be sufficiently small to keep $\mathbf{x}$ within one orthant; we simply do not consider $\mathbf{x}$-vectors not in the proper orthant.

2. Application of the method to ill-posed problems. The extension of the above described linear programming technique to the problem of computing confidence interval bounds has wide application in solving problems arising in physical situations modeled by Fredholm integral equations of the first kind:

$$\int_a^b K(t,s)x(s)\,ds = y(t), \tag{2.1}$$

where $K(t,s)$ and $y(t)$ are known functions and $x(s)$ is the unknown to be estimated. In practice one does not usually know $y(t)$ exactly but rather has a measured pointwise approximation so that (2.1) is replaced by the system

$$\int_a^b K_i(s)x(s)\,ds = \hat{y}_i + \hat{\varepsilon}_i, \qquad 1,2,\cdots,m, \tag{2.2}$$

where the $\hat{\varepsilon}_i$ are unknown stochastic measuring errors which are assumed to have mean $\mathbf{0}$ and a known covariance matrix $\mathbf{S}^2$, i.e.,

$$E(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}, \qquad E(\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}^T) = \mathbf{S}^2. \tag{2.3}$$

There is no loss of generality in assuming that $\mathbf{S}^2$ is a diagonal matrix. The functions $K_i(s)$ are generally not known exactly either, but it is often possible to measure them quite accurately relative to the accuracy obtainable for the $\hat{y}_i$. We assume that we can determine pointwise approximations $K_i(s_j)$ at as fine a mesh s_j as may be necessary to give a discrete problem that accurately models the system of continuous equations. The discrete problem obtained by applying a quadrature rule to (2.2) can be written

$$\mathbf{x} = \hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}} \tag{2.4}$$

where $\mathbf{K}$ is an $m \times n$ matrix composed of the $K_i(s_j)$ and the quadrature weighting coefficients and $\mathbf{x}$ is an n-vector whose elements constitute a pointwise approximation to $\mathbf{x}(s)$. We assume that there is essentially no uncertainty in $\mathbf{K}$.

It is important to take n large enough so that the discretization errors are negligible in comparison to the statistical errors $\hat{\varepsilon}_i$. If n is not chosen large enough so that (2.4) accurately models the physical situation, then because the problem is ill-conditioned, the resulting solution set may not be consistent with physically motivated a priori constraints. The most commonly encountered a priori physical constraint is that the x_j be nonnegative. One of the first studies of the use of nonnegativity constraints in problems of this type was that of W. R. Burrus [1] who introduced the idea of constrained interval estimation. Burrus extended the classical statistical interval estimation technique to take into account the a priori constraints in order to reduce the sizes of the confidence intervals obtained (cf. also Rust and Burrus [6] and Replogle, Holcombe and Burrus [7]).

In order to use the linear programming techniques described in § 1 to construct confidence interval estimates for the x_j, it is necessary to know the probability distributions of the errors $\hat{\varepsilon}_i$. For physical problems, it is usually assumed that the $\hat{\varepsilon}_i$ are independently normally distributed with mean $\mathbf{0}$ and covariance matrix $\mathbf{S}^2 = \operatorname{diag}(s_1^2, \cdots, s_m^2)$. These assumptions imply that $y_i \sim N(\bar{y}_i, s_i)$, hence $(y_i - \bar{y}_i)/s_i \sim N(0, 1)$, where we denote by $\bar{\mathbf{y}}$ the "true" vector $\mathbf{y}$ that would be obtained if there were no measuring errors, i.e., the unknown mean vector of the $\mathbf{y}$-distribution. For any constant μ define a box in $\mathbf{y}$-space by

$$\mathcal{B}_{\hat{y}}(\mu) = \left\{ \mathbf{y} \,\middle|\, \max_i \frac{|y_i - \hat{y}_i|}{s_i} \leqq \mu \right\}.$$

Given a probability level α, we wish to choose μ so that $\Pr\{\bar{\mathbf{y}} \in \mathcal{B}_{\hat{y}}(\mu)\} \geqq \alpha$.

$$\begin{aligned}\Pr\{\bar{\mathbf{y}} \in \mathcal{B}_{\hat{y}}(\mu)\} &= \Pr\left\{ \max_i \left(\frac{|\bar{y}_i - \hat{y}_i|}{s_i} \right) \leqq \mu \right\} \\ &= \Pr\left\{ \left(\frac{|\bar{y}_1 - \hat{y}_1|}{s_1} \leqq \mu \right) \wedge \left(\frac{|\bar{y}_2 - \hat{y}_2|}{s_2} \leqq \mu \right) \wedge \cdots \wedge \left(\frac{|\bar{y}_m - \hat{y}_m|}{s_m} \leqq \mu \right) \right\} \\ &= \prod_{i=1}^{m} \Pr\{\hat{y}_i - \mu s_i \leqq \bar{y}_i \leqq \hat{y}_i + \mu s_i\}.\end{aligned}$$

The last equality follows because the y_i are independently distributed. Now let $\Pr\{\hat{y}_i - \mu s_i \leqq \bar{y}_i \leqq \hat{y}_i + \mu s_i\} \geqq \beta$, $i = 1, 2, \cdots, m$. Then $\Pr\{\bar{\mathbf{y}} \in \mathcal{B}_{\hat{y}}(\mu)\} \geqq \beta^m$. Hence, μ must be chosen such that $\beta \geqq \alpha^{1/m}$, i.e., μ must satisfy

$$\frac{1}{\sqrt{2\pi}} \int_{-\mu}^{\mu} \exp(-\eta^2/2)\, d\eta = \alpha^{1/m},$$

and hence can be obtained from standard tables of the normal distribution. In this case the Δy_i of § 1 are equal to μs_i, $i = 1, 2, \cdots, m$. Using this $\mathbf{\Delta y}$ and the knowledge of the orthant of $\mathbf{x}$, we apply the methods of § 1 to obtain a confidence region for $\mathbf{x}$. The required tableau is

$$\begin{bmatrix} \mathbf{K} \\ -\mathbf{K} \end{bmatrix} \mathbf{x} \leqq \begin{bmatrix} \hat{\mathbf{y}} + \mu \mathbf{S}\mathbf{e} \\ -\hat{\mathbf{y}} + \mu \mathbf{S}\mathbf{e} \end{bmatrix} \tag{2.5}$$

where $\mathbf{e} = (1, 1, 1, \cdots, 1)^T$. This tableau is a special case of (1.7) in which $\mathbf{S}_g = \mathbf{I}$ and $\Delta \mathbf{A} = 0$.

Another approach to the problem defined by (2.3) and (2.4) is to notice that it is formally identical to the statement of the classical linear regression model. To construct a confidence region for $\mathbf{x}$, we choose a confidence level α for $\mathbf{y}$ and consider the corresponding confidence ellipsoid in $\mathbf{y}$-space:

$$(\mathbf{y} - \hat{\mathbf{y}})^T S^{-2} (\mathbf{y} - \hat{\mathbf{y}}) \leqq \mu^2 \tag{2.6}$$

where μ is determined by α. If we assume that $\mathbf{y}$ has an m-dimensional multivariate normal distribution with mean $\hat{\mathbf{y}}$ and variance–covariance matrix $\mathbf{S}^2$, it follows that $(\mathbf{y} - \hat{\mathbf{y}})^T S^{-2} (\mathbf{y} - \hat{\mathbf{y}})$ has a chi-squared distribution with m degrees of freedom (cf. [8, Appendix V]). The relationship between the probability level α and the parameter μ^2 is then

$$\alpha = \int_0^{\mu^2} \frac{\rho^{m/2-1} \exp(-\rho/2)}{\Gamma(m/2)2^{m/2}}\, d\rho \tag{2.7}$$

where Γ denotes the gamma function. This relationship is tabulated in standard tables of the χ^2-distribution.

In order to relate the ellipsoid defined by (2.6) to the physical problem it is helpful to consider an underlying system of equations

$$\mathbf{K}\bar{\mathbf{x}} = \bar{\mathbf{y}} \tag{2.8}$$

where $\bar{\mathbf{x}}$ is an unknown, in some sense "true" approximation to $\mathbf{x}(s)$, and $\bar{\mathbf{y}}$ is the corresponding, unknown vector in the observation space. The assumption here is that n has been chosen large enough so that $(\mathbf{K}\bar{\mathbf{x}})_i$ gives a very accurate representation of $\int_a^b K_i(s)\mathbf{x}(s)\, ds$ and that $\bar{\mathbf{y}}$ is the vector of observations that would be obtained if it were possible to completely eliminate stochastic measuring errors. The sample vector $\hat{\mathbf{y}}$, which contains measuring errors, is a point estimate of $\bar{\mathbf{y}}$, and the ellipsoid (2.6) is a confidence ellipsoid for $\bar{\mathbf{y}}$, i.e.,

$$\Pr\{(\bar{\mathbf{y}}-\hat{\mathbf{y}})^T\mathbf{S}^{-2}(\bar{\mathbf{y}}-\hat{\mathbf{y}}) \leqq \mu^2\} \geqq \alpha. \tag{2.9}$$

The problem of finding confidence intervals for the x_j can be formulated as follows: taking $\mathbf{e}_j$ to be the n-vector with 1 in the jth place and 0 elsewhere the confidence bounds are given by

$$\begin{aligned} x_j^{lo} &= \min_{\mathbf{x}} \{\mathbf{e}_j^T\mathbf{x} | \mathbf{x} \geqq 0 \text{ and } (\mathbf{K}\mathbf{x}-\hat{\mathbf{y}})^T\mathbf{S}^{-2}(\mathbf{K}\mathbf{x}-\hat{\mathbf{y}}) \leqq \mu^2\}, \\ x_j^{hi} &= \max_{\mathbf{x}} \{\mathbf{e}_j^T\mathbf{x} | \mathbf{x} \geqq 0 \text{ and } (\mathbf{K}\mathbf{x}-\hat{\mathbf{y}})^T\mathbf{S}^{-2}(\mathbf{K}\mathbf{x}-\hat{\mathbf{y}}) \leqq \mu^2\} \end{aligned} \tag{2.10}$$

where $\mathbf{e}_j^T\mathbf{x}$ is the objective function for constrained optimization, the constraints being $\mathbf{x} \geqq 0$ and $(\mathbf{K}\mathbf{x}-\hat{\mathbf{y}})^T\mathbf{S}^{-2}(\mathbf{K}\mathbf{x}-\hat{\mathbf{y}}) \leqq \mu^2$. More generally, confidence interval estimates for any linear function

$$\varphi(x) = \mathbf{w}^T\mathbf{x} \tag{2.11}$$

of the x_j can be obtained from

$$\begin{aligned} \varphi^{lo} &= \min_{\mathbf{x}} \{\mathbf{w}^T\mathbf{x} | \mathbf{x} \geqq 0 \text{ and } (\mathbf{K}\mathbf{x}-\hat{\mathbf{y}})^T\mathbf{S}^{-2}(\mathbf{K}\mathbf{x}-\hat{\mathbf{y}}) \leqq \mu^2\}, \\ \varphi^{hi} &= \max_{\mathbf{x}} \{\mathbf{w}^T\mathbf{x} | \mathbf{x} \geqq 0 \text{ and } (\mathbf{K}\mathbf{x}-\hat{\mathbf{y}})^T\mathbf{S}^{-2}(\mathbf{K}\mathbf{x}-\hat{\mathbf{y}}) \leqq \mu^2\}. \end{aligned} \tag{2.12}$$

Such estimates are useful in physical problems when it is desired to determine integral quantities of the form

$$\varphi[x(s)] = \int_a^b w(s)x(s)\, ds. \tag{2.13}$$

The form (2.11) can be obtained from (2.13) by applying the same quadrature rule used to reduce (2.2) to (2.4).

In theory there are a number of ways to calculate solutions to the problems (2.10) and (2.12). The most obvious approach is to apply the Wolfe duality theorem to obtain a linearly constrained quadratic maximization problem. A similar approach involving parametric quadratic programming has been described in [6, Chap. 5]. In practice these approaches have so far not been successful. We believe that this failure is caused by the ill-conditioning of the problem. The failure of standard optimization codes to solve these ill-conditioned problems has led us to consider suboptimal approximating problems which are computationally more tractable. These are obtained by replacing the y-ellipsoid by a circumscribing polytope (cf. [9, Chap. 3]) to obtain linear programming problems which yield confidence interval estimates that are wider than the optimal one that would result from solving the quadratic problem. It is easy to see that the ∞-norm polytope yields a tableau formally identical to (2.5), the only difference being that μ is chosen by means of the χ^2 rather than the normal distribution so that the intervals obtained are wider than in the previous case. An alternative is to use a 1-norm circumscribing polytope, which has fewer corners and hence fewer potential extreme points than the ∞-norm box. In some cases this gives narrower interval estimates, but the 1-norm box gives rise to a $(2m+1)\times(m+n)$ linear programming problem, whereas the size of (2.5) is only $2m\times n$. Intersecting the two circumscribing boxes should give better results than either alone, but the tableau is enlarged to $(3m+1)\times(m+n)$. Experience has shown that with ill-conditioned problems which are large to begin with, enlarging the tableau will often compound numerical difficulties, so that the possible advantages of intersecting the boxes are lost. It is an open question whether the approach to take is problem-dependent.

3. Examples. We conclude by presenting three examples, first the 4×4 example of Oettli [3], then a 5×5 Hilbert example, and finally, a physical problem involving the Fredholm integral equation already mentioned. All of the solutions were obtained using either the tableau (1.7)′ or (2.5).

Example 1. Oettli's Problem.

$$\mathbf{A}=\begin{bmatrix} 4.33 & -1.12 & -1.08 & 1.14 \\ -1.12 & 4.33 & .24 & -1.22 \\ -1.08 & .24 & 7.21 & -3.22 \\ 1.14 & -1.22 & -3.22 & 5.43 \end{bmatrix},$$

$$\mathbf{b}=\begin{bmatrix} 3.52 \\ 1.57 \\ .54 \\ -1.09 \end{bmatrix}, \qquad \Delta a_{ij}=\Delta b_i=.005.$$

In this problem, Δa_{ij} and Δb_i are small enough so that all solutions lie in the same orthant. The true solution vector is

$$\mathbf{x}=\begin{bmatrix} 1.0462 \\ .5627 \\ .1110 \\ -.2281 \end{bmatrix}.$$

The bounds obtained using the tableau (1.7)′ are the same as those of [3], i.e.,

$$\begin{bmatrix} 1.040 \\ .556 \\ .105 \\ -.236 \end{bmatrix} \leqq \mathbf{x} \leqq \begin{bmatrix} 1.052 \\ .569 \\ .117 \\ -.221 \end{bmatrix}.$$

Figure 1 is a plot of $x(3)$ vs. $x(1)$ for several hundred solutions of the Oettli problem as **A** and **b** are varied within their bounds. Note that the spread in each component of the solution vector is on the order of .01, about the same as the uncertainty in the elements of **A** and **b**. Thus the Oettli problem is a very well-conditioned one.

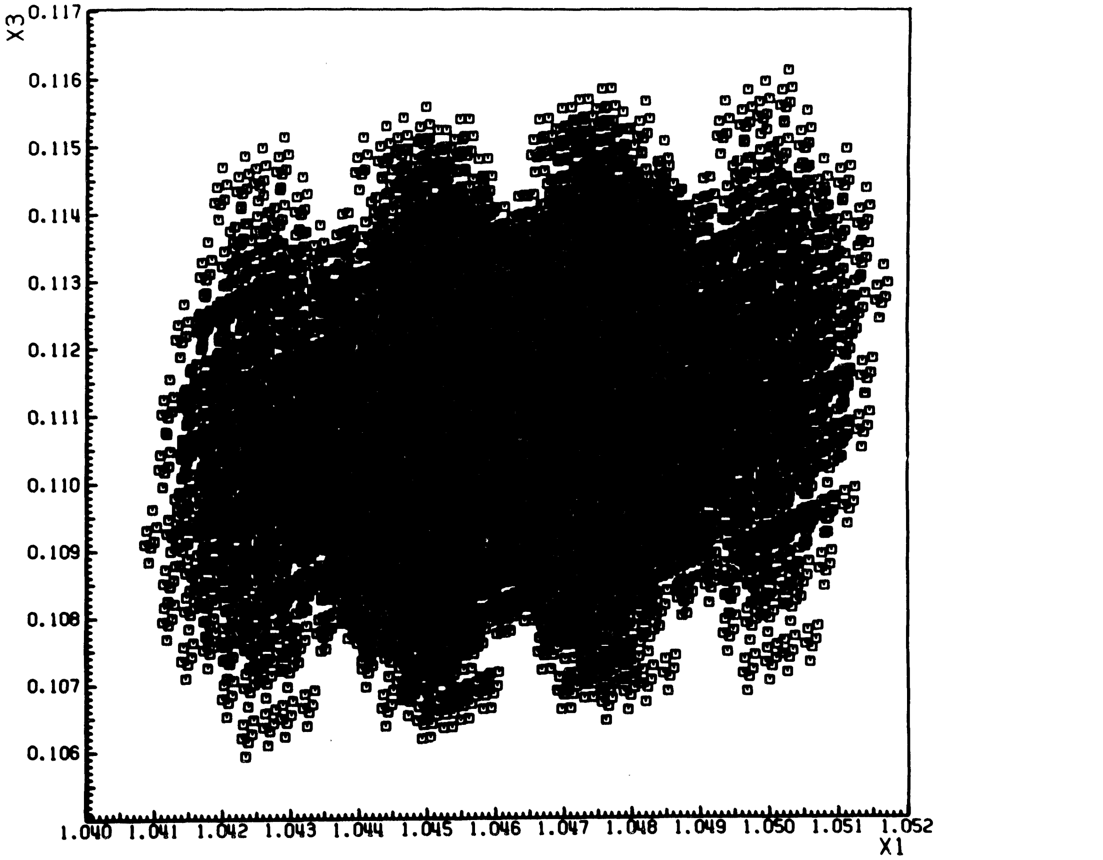


FIG. 1

Example 2. A Hilbert Problem. The 5×5 Hilbert example is an ill-conditioned problem which shows how the orthant constraints can be helpful in reducing the size of the allowed solution set. To form the 5×5 Hilbert matrix **A**, set $a_{ij}=1/(i+j-1)$. Let

$$\mathbf{b}^0=\begin{bmatrix}.2837\times 10^5\\ .2210\times 10^5\\ .1817\times 10^5\\ .1545\times 10^5\\ .1344\times 10^5\end{bmatrix},\qquad \mathbf{b}^1=\begin{bmatrix}.2837\times 10^5\\ .2210\times 10^5\\ .1818\times 10^5\\ .1545\times 10^5\\ .1345\times 10^5\end{bmatrix}.$$

Suppose that the true right-hand side is $\mathbf{b}^0$, that we know **A** exactly, and that $\Delta b_i=10$. Then $\mathbf{b}^1$ falls within the range $\mathbf{b}^0\pm\Delta\mathbf{b}$. The exact solution to the system

$\mathbf{Ax}=\mathbf{b}^0$ is

$$\mathbf{x}^0=\begin{bmatrix} .20\times 10^3 \\ .30\times 10^4 \\ .32\times 10^5 \\ .40\times 10^5 \\ .30\times 10^5 \end{bmatrix}.$$

The solution to the system $\mathbf{Ax}=\mathbf{b}^1$, correctly rounded to four places, is

$$\mathbf{x}^1=\begin{bmatrix} .1113\times 10^4 \\ -.1316\times 10^4 \\ .9823\times 10^5 \\ -.5603\times 10^5 \\ .7552\times 10^5 \end{bmatrix}.$$

In a physical problem involving measurement of the right-hand side, it might be expected to have errors in the fourth place. $\mathbf{x}^1$ would be a meaningless solution to a physical system where the solution vector is known to lie in the positive orthant. Putting positive orthant bounds on $\mathbf{x}$, however, we obtain the following reasonable bounds on the solutions of the two systems $\mathbf{Ax}=\mathbf{b}^0$, and $\mathbf{Ax}=\mathbf{b}^1$:

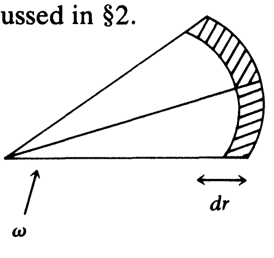
$$\begin{bmatrix} 0. \\ 0. \\ 0. \\ .168\times 10^4 \\ 0. \end{bmatrix} \leqq \mathbf{x}^0 \leqq \begin{bmatrix} .5075\times 10^3 \\ .9099\times 10^4 \\ .5237\times 10^5 \\ .9588\times 10^5 \\ .5142\times 10^5 \end{bmatrix},$$

$$\begin{bmatrix} 0. \\ 0. \\ 0. \\ .1338\times 10^5 \\ 0. \end{bmatrix} \leqq \mathbf{x}^1 \leqq \begin{bmatrix} .5788\times 10^3 \\ .8986\times 10^4 \\ .4790\times 10^5 \\ .9591\times 10^5 \\ .4392\times 10^5 \end{bmatrix}.$$

In each case, the bounds include the vector $\mathbf{x}^0$, the solution to the true system $\mathbf{Ax}=\mathbf{b}^0$. Figures 2a and 2b show plots obtained by Monte Carlo sampling of $x(4)$ vs. $x(3)$ for a large number of solutions for sample problems with $\Delta b_i \leqq 10$. Figure 2a represents all solutions outside the positive orthant, while Fig. 2b represents all solutions in the positive orthant. The graphs show that the solutions are very sensitive to small changes in $\mathbf{b}$, and would be practically unbounded if it were not for the orthant constraint. Since many physical problems are ill-conditioned, the Hilbert example demonstrates the usefulness of an orthant constraint in providing meaningful solutions to problems where there may be errors in input data.

Example 3. A stellar density problem. We now present an example of the Fredholm integral equation discussed in §2.

(3.1)



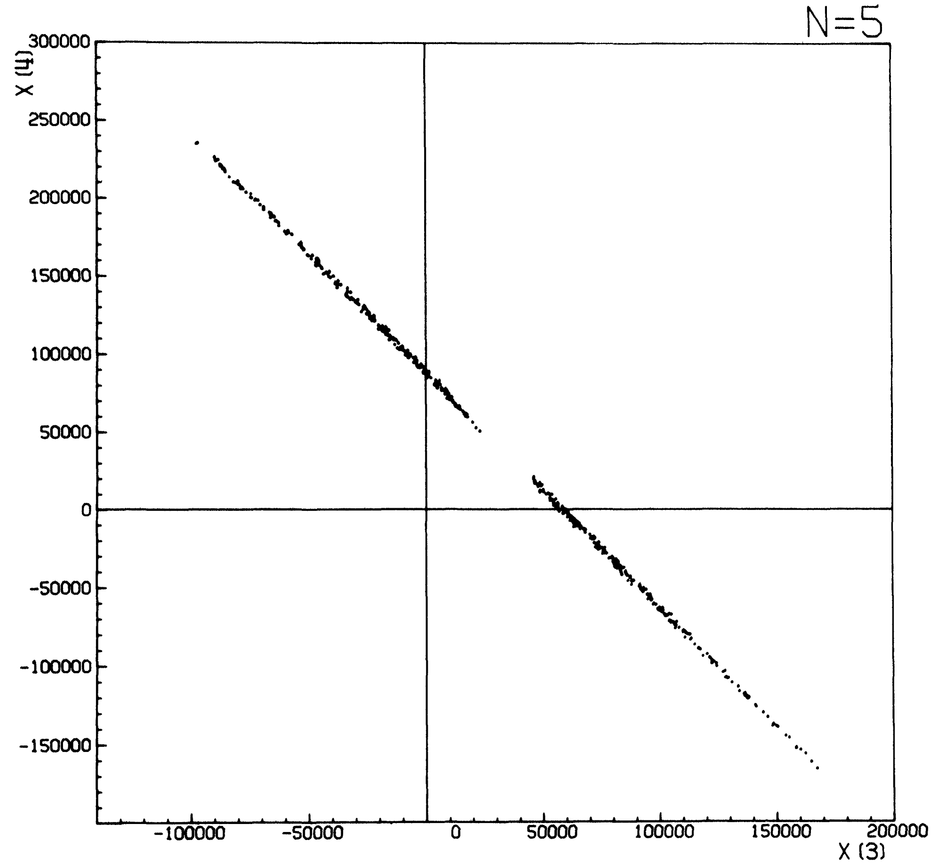


FIG. 2a

The problem is to estimate density of stars of a given type in a volume $dV = \omega r^2 dr$, where ω is the solid angle covered by the observations. Images of stars on a photographic plate are classified by brightness (magnitude) m. The integral equation to be solved, for density $D(r)$, is:

$$\int_m c(m)\,dm = \int_r \int_m [\Phi(m+5-5\log r - a(r))]\,dm\,\omega r^2 D(r)\,dr \tag{3.2}$$

where

r = distance from the sun, measured in parsecs (1 parsec = 3.25 light years),
ω = solid angle in steradians,
$c(m)$ = number of stars of apparent magnitude m,
$a(r)$ = interstellar absorption in magnitudes,
M = absolute magnitude $= m+5-5\log r - a(r)$, a measure of intrinsic luminosity,
$\Phi(M) = \dfrac{1}{\sqrt{2\pi\sigma}} \exp\{-[\frac{1}{2}(M-Mo)^2/\sigma]\}$ = luminosity function,
M_0 = mean absolute magnitude for the stars in question,
σ^2 = dispersion in absolute magnitude.

It is clear from the expression for $\Phi(M)$ that we are assuming a Gaussian distribution for the intrinsic luminosities of the stars in question.

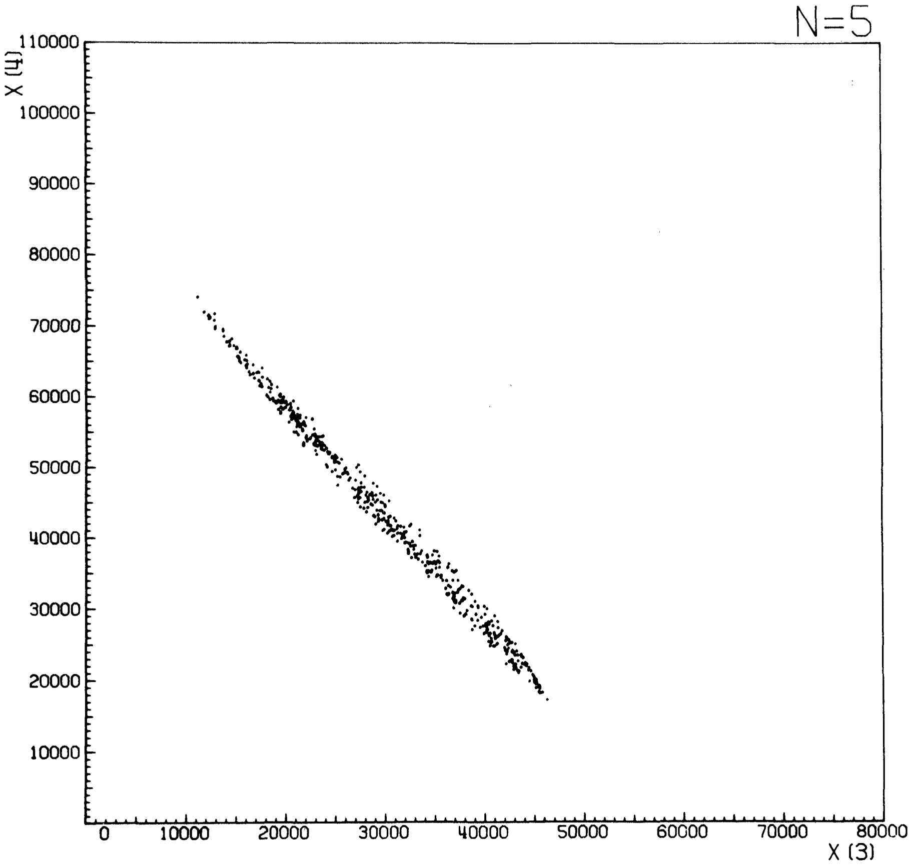


FIG. 2b

We put (3.2) in the form more usually seen by discretizing m. Stars are placed into magnitude bins when counted. Let $c(m_i)$ be the number of stars between magnitude m_i and m_{i+1}. We then have

$$c(m_i)=\int_{r_{\min}}^{r_{\max}}\left[\int_{m_i}^{m_{i+1}}\Phi(m+5-5\log r-a(r))\,dm\right]\omega r^2D(r)\,dr, \tag{3.3}$$

where $r_{\min}$ and $r_{\max}$ are determined by the brightest star observed on the plate and the plate limit. Equation (3.3) is a Fredholm integral equation of the first kind with kernel $K(i, r)=\int_{m_i}^{m_{i+1}}\Phi(m+5-5\log r-a(r))\,dm$ which can easily be evaluated by numerical quadrature for any r, m_i, and m_{i+1}. $c(m_i)$ is observed by counting images of stars on a photographic plate. We assume that the counting errors are Poisson distributed so that the square root error law applies.

To solve (3.3) we now discretize r and apply a quadrature rule for integration.

$$c(m_i)\approx\sum_j K(i, r_j)w_j\omega r_j^2D(r_j) \tag{3.4}$$

where w_j is the weight associated with the quadrature rule that is used. In this case a simple rectangular rule was used. We can now put (3.4) in the form of a system of linear equations. To form the matrix **A**, we set $a_{ij}=K(i, r_j)w_j\omega r_j^2$. Note that the *columns* of **A** correspond to r-mesh points and the *rows* of $\mathbf{A}$ to magnitude bins. For the right-hand side, $b_i = c(m_i)$, $\Delta b_i = \sqrt{c(m_i)}$, and we solve for the vector $D(r_j)$. We choose the magnitude bins to contain enough stars so that the Poisson distribution assumed in estimating the counting error is reasonably approximated by a normal distribution which is used in making confidence interval statements about the solution bounds. Choosing $\Delta b_i = \sqrt{c(m_i)}$ means that we are seeking the 66.7% confidence interval.

In general, $\mathbf{A}$ has many more columns than rows. As an example, we choose G8–G9 stars of luminosity, class III [2]. In this case,

$$\begin{aligned} \sigma &= .8 & \omega &= .1206 \text{ square radians,} \\ M_0 &= 1, & r_{\min} &= 75.\text{ p.c.,} \\ a(r) &= 0, & r_{\max} &= 7000 \text{ p.c..} \end{aligned}$$

In order to have a problem whose solution vector is known, we used the above data to generate $\mathbf{A}$, then put in a simulated solution generated by

$$D(r_j) = \exp\left[\frac{(-\ln(r_j - 75.) - \ln 75.)^2}{2\sigma_1^2}\right]$$

with $\sigma_1^2 = 2$.

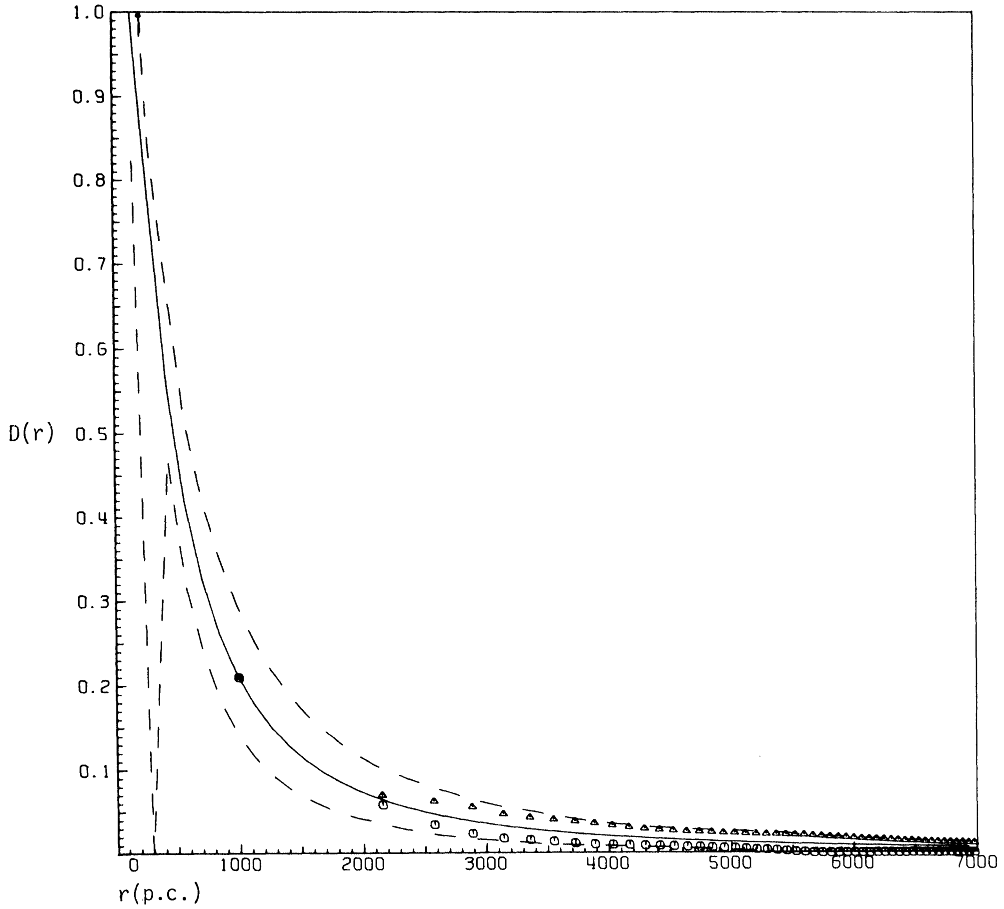


FIG. 3

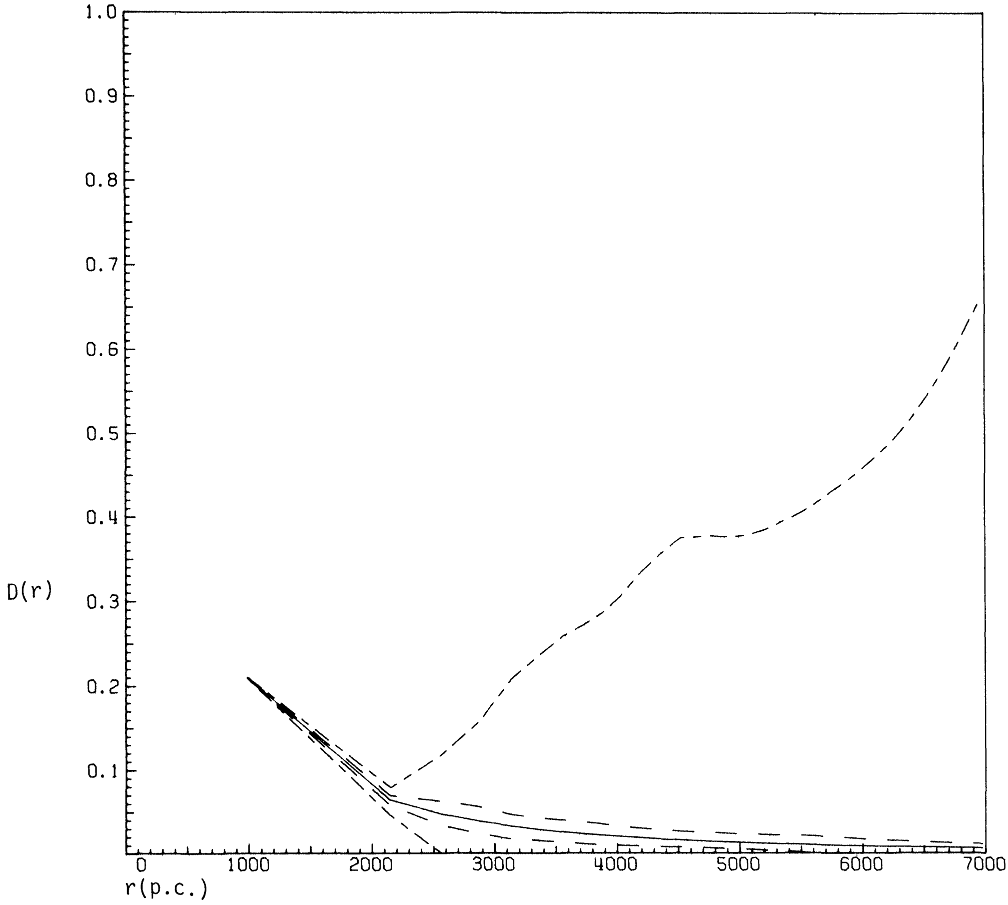


FIG. 4

We then multiplied **A** by **D** to create a right-hand side. $D(r)$ is a plausible density function for this problem. Further, $D(r)$ is always nonnegative and is a monotonic nonincreasing function of r for $r \geqq 75$. We chose a monotonic solution in order to simulate a problem obtained from plates covering an area of the sky in the direction of the north galactic pole, i.e. "looking" in a direction perpendicular to the plane of the galaxy. We used 13 magnitude bins and 50 r-mesh points. Two different spacings of mesh points were tried; first, an equally spaced r-mesh was used, and then the mesh points were spaced so as to yield equal volumes dV. Trials were made with both mesh spacings, with and without the monotonicity constraint. Figure 3 is a graph of the true solution and bounds obtained using nonnegativity and the monotonicity constraint. The solid line is the true solution, the dashed lines represent the bounds obtained using equal r-mesh spacing and the symbols are the bounds obtained using equal volume mesh. Figure 4 gives the results using equal volume spacing both with and without the monotonicity constraint.

Acknowledgments. We would like to thank Walter Burrus for his generous advice, Joze Rant for his work in improving the linear programming algorithm and Collene Trammell for her cheerful cooperation and careful work in typing this manuscript.

REFERENCES

[1] W. R. BURRUS, *Utilization of A Priori Information by Means of Mathematical Programming in the Statistical Interpretation of Measured Distributions*, ORNL-3743, June, 1965.

[2] A. R. UPGREN, JR., *The space distribution of late-type stars in a north galactic pole region*, Astron. J., 67 (1962), no. 1 pp. 37–78.

[3] W. OETTLI, *On the solution set of a linear system with inaccurate coefficients*, this Journal, 2 (1965), pp. 115–18.

[4] W. OETTLI AND W. PRAGER, *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides*, Numer. Math., 6 (1964), pp. 405–09.

[5] W. OETTLI, W. PRAGER AND J. H. WILKINSON, *Admissible solutions of linear systems with not sharply defined coefficients*, this Journal, 2 (1965), pp. 291–99.

[6] B. W. RUST AND W. R. BURRUS, *Mathematical Programming and the Numerical Solution of Linear Equations*, American Elsevier, New York, 1972.

[7] J. REPLOGLE, B. D. HOLCOMBE AND W. R. BURRUS, *The use of mathematical programming for solving singular and poorly conditioned systems of equations*, J. Math. Anal. and Appl., 20 (1967), pp. 310–24.

[8] H. SCHEFFÉ, *The Analysis of Variance*, John Wiley, New York, 1967.

[9] B. W. RUST AND W. R. BURRUS, *Suboptimal Methods for Solving Constrained Estimation Problems*, DASA 2604, January, 1971.