


Confidence Intervals for Discrete Approximations to Ill-Posed Problems
Author(s): Bert W. Rust and Dianne P. O'Leary
Source: *Journal of Computational and Graphical Statistics*, Vol. 3, No. 1 (Mar., 1994), pp. 67-96
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of America
Stable URL: https://www.jstor.org/stable/1390796
Accessed: 29-04-2020 15:19 UTC

Confidence Intervals for Discrete Approximations to Ill-Posed Problems

Bert W. RUST* and Dianne P. O'LEARY†

We consider the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ that is obtained by discretizing a system of first-kind integral equations describing a set of physical measurements. The n vector $\boldsymbol{\beta}$ represents the desired quantities, the $m \times n$ matrix $\boldsymbol{X}$ represents the instrument response functions, and the m vector $\boldsymbol{Y}$ contains the measurements actually obtained. These measurements are corrupted by random measuring errors $\boldsymbol{\epsilon}$ drawn from a distribution with zero mean vector and known variance matrix. Solution of first-kind integral equations is an ill-posed problem, so the least squares solution for the above model is a highly unstable function of the measurements, and the classical confidence intervals for the solution are too wide to be useful. The solution can often be stabilized by imposing physically motivated nonnegativity constraints. In a previous article (O'Leary and Rust 1986) we developed a method for computing sets of nonnegatively constrained simultaneous confidence intervals. In this article we briefly review the simultaneous intervals and then show how to compute nonnegativity constrained one-at-a-time confidence intervals. The technique gives valid confidence intervals even for problems with $m < n$. We demonstrate the methods using both an overdetermined and an underdetermined problem obtained by discretizing an equation of Phillips (Phillips 1962).

Key Words: Confidence intervals; First-kind integral equations; Ill-posed problems; Linear regression; Nonnegatively constrained regression.

1. INTRODUCTION

A central problem in statistics is the estimation of the quantities

$$\phi_k^* = \boldsymbol{w}_k^T \boldsymbol{\beta}, \quad k = 1, 2, \ldots, p, \tag{1.1}$$

where the $\boldsymbol{w}_k$ are given n vectors, and $\boldsymbol{\beta}$ satisfies the linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathcal{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}, \quad \mathcal{E}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right) = \Sigma^2, \tag{1.2}$$

with $\boldsymbol{Y}$ a measured m vector, $\boldsymbol{X}$ a known $m \times n$ matrix, $\boldsymbol{\beta}$ an unknown n vector, and Σ^2 a known positive definite variance matrix. We focus in this article on finding

*Mathematician, Applied and Computational Mathematics Division, National Institute of Standards and Technology, Gaithersburg, MD 20899

†Mathematician, Applied and Computational Mathematics Division, National Institute of Standards and Technology, Gaithersburg, MD 20899

Journal of Computational and Graphical Statistics, Volume 3, Number 1, Pages 67–96

confidence intervals for the ϕ_k^* when it is known that $\boldsymbol{\beta}$ is nonnegative. Such nonnegativity constrained problems are widespread, arising whenever the parameters represent physical quantities such as counts, densities, or pressures. They often arise from discretizing a system of first-kind integral equations

$$Y_i = \int_a^b X_i(\xi)\beta(\xi)\,d\xi + \epsilon_i, \quad i = 1, \ldots, m, \tag{1.3}$$

where $\beta(\xi)$ is an unknown, nonnegative function, the $X_i(\xi)$ are known response functions for a measuring instrument, the Y_i are the discrete measurements, and the ϵ_i are random measuring errors. Such integral equations are fundamental for the reconstruction of signals from noisy data, with applications in medical imaging, remote sensing of global climate data, seismic exploration, and many other areas.

Without severe a priori restrictions on $\beta(\xi)$, the m discrete relationships (1.3) cannot determine the value of $\beta(\xi)$ at every point on the interval $[a, b]$. Therefore, we seek instead to estimate a finite set of average values of $\beta(\xi)$, averaged on various subintervals of $[a, b]$. More precisely, we choose p representative points $\xi_1 < \xi_2 < \cdots < \xi_p$ in $[a, b]$ and for each ξ_k choose a corresponding window function $w_k(\xi)$ designed so that

$$\phi_k^*(\xi_k) = \int_a^b w_k(\xi)\beta(\xi)\,d\xi \tag{1.4}$$

is an average value of the unknown function in a subinterval around ξ_k. Discretizing these averaging integrals by the same method used to discretize (1.3) gives a set of p linear functions (1.1) that approximate the desired averages. Note that the individual components of $\boldsymbol{\beta}$ can be specified by choosing $p = n$ and $\boldsymbol{w}_k = \boldsymbol{e}_k$, $k = 1, 2, \ldots, n$, where $\boldsymbol{e}_k$ is the kth column of the identity matrix.

Having accomplished the discretization, the task becomes one of estimating each of the p linear functions (1.1) subject to the statistical constraints (1.2). These estimates can be either point estimates or confidence interval estimates. Following Graybill (1976) we consider two kinds of confidence intervals:

1. **One-at-a-time confidence intervals:** Each ϕ_k^* is treated individually, and for a prespecified probability α ($0 < \alpha < 1$), the $100\alpha\%$ confidence intervals are determined separately. Each individual interval contains the corresponding true value with probability α—that is,

$$\Pr\left\{ \phi_k^{\text{lo}} \le \boldsymbol{w}_k^T\boldsymbol{\beta} \le \phi_k^{\text{up}} \right\} = \alpha, \quad k = 1, 2, \ldots, p. \tag{1.5}$$

2. **Simultaneous confidence intervals:** All of the ϕ_k^* are treated together, and the intervals are constructed so that, with probability α, they all contain the corresponding true values—that is,

$$\Pr\left\{\phi_k^{\text{lo}} \le \boldsymbol{w}_k^T\boldsymbol{\beta} \le \phi_k^{\text{up}}, \quad k = 1, 2, \ldots, p \right\} \ge \alpha. \tag{1.6}$$

Clearly this is a more exacting requirement than that for one-at-a-time intervals, and thus for a given α each of the simultaneous intervals is wider than its one-at-a-time counterpart.

The inversion of first-kind integral equations is an ill-posed problem, so the matrix $\boldsymbol{X}$ is usually ill-conditioned or even rank deficient. As a result the confidence intervals are typically extremely wide or even unbounded. To get physically plausible solutions, it is necessary to impose a priori constraints (e.g., nonnegativity) on the solution. Good algorithms for computing nonnegatively constrained point estimates have been widely available for some time (Lawson and Hanson 1974). In previous work (O'Leary and Rust 1986) we developed an algorithm for computing nonnegatively constrained, simultaneous confidence intervals when ϵ is normally distributed. In this article we extend these techniques to include nonnegatively constrained one-at-a-time intervals.

In Section 2 we present two example problems used to illustrate our results. In Section 3 we briefly review classical point-estimation theory with an emphasis on the effects of ill conditioning and rank deficiency. In Section 4 we review classical, one-at-a-time confidence intervals, and in Section 5 we consider the geometrical basis for interval estimation, characterizing the bounds as solutions to certain constrained optimization problems that will later be augmented to include the nonnegativity constraints. In Section 6 we review the estimation of classical, simultaneous confidence intervals whose geometrical basis is identical to that for the one-at-a-time intervals. The purpose of the discussion in these initial sections is to note the deficiencies of the classical bounds and point estimates when used for discretizations of ill-posed problems. In Section 7 we formally append the constraints $\boldsymbol{\beta} \geq \mathbf{0}$ to the classical linear regression model and demonstrate their effectiveness in stabilizing the point estimates. We also show that these constraints almost always produce bounded confidence intervals, even for underdetermined problems, and show how to extend the theory for simultaneous intervals to accommodate them. The extension of the classical theory for one-at-a-time intervals is considerably more difficult, but in Section 8 we prove a conjecture, first made almost 30 years ago by Walter R. Burrus (1965), that allows the calculation of such intervals using the same algorithm we originally developed for simultaneous intervals.

Because we assume that the only significant errors are the random perturbations ϵ in the data, we implicitly assume that the model described by (1.1) and (1.2) is correct. In particular, this implies that errors due to discretization of the integrals are negligible. Thus we produce confidence intervals corresponding to the class of functions for which the quadrature is sufficiently accurate (or the instrument resolution sufficiently fine).

2. TEST PROBLEMS

To generate problems with known solutions, we use the integral equation studied by Phillips (1962),

$$Y(t) = \int_{-3}^{3} X(t,\xi)\beta(\xi)\, d\xi, \quad -6 \leq t \leq 6, \tag{2.1}$$

where

$$\begin{aligned} X(t,\xi) &= 1 + \cos\left[\tfrac{\pi(\xi - t)}{3}\right], \quad |\, \xi - t \,| \leq 3, |\, t \,| \leq 6 \\ &= 0, \qquad\qquad\qquad\quad \text{otherwise}, \end{aligned} \tag{2.2}$$

and

$$\begin{aligned} Y(t) &= (6-\mid t \mid)\left[1+\tfrac{1}{2}\cos\left(\tfrac{\pi t}{3}\right)\right]+\tfrac{9}{2\pi}\sin\left(\tfrac{\pi|t|}{3}\right), \quad \mid t \mid \leq 6 \\ &= 0, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{aligned} \tag{2.3}$$

The kernel $X(t,\xi)$ is nonnegative, a property common in many real-world applications. The function $Y(t)$ is also nonnegative, symmetric about $t=0$, and bell shaped, with maximum value $Y(0)=9$ and minimum value $Y(-6)=0=Y(+6)$. The exact solution is

$$\begin{aligned} \beta(\xi) &= 1+\cos\left(\tfrac{\pi\xi}{3}\right), \quad \mid \xi \mid \leq 3 \\ &= 0, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{aligned} \tag{2.4}$$

This nonnegative function also defines a symmetric bell-shaped curve with a maximum $\beta(0)=2$ and minima $\beta(-3)=0=\beta(+3)$.

To get a system of integral equations, we chose m equally spaced mesh points t_i ranging from $t_1=-5.925$ to $t_m=+5.925$ and defined $X_i(\xi)\equiv X(t_i,\xi)$. This gave

$$Y_i \equiv Y(t_i)=\int_{-3}^{3} X_i(\xi)\beta(\xi)\,d\xi, \quad i=1,2,\ldots,m. \tag{2.5}$$

We used the trapezoidal rule with mesh spacing $\Delta\xi=.05$ to reduce this system of integral equations to a linear algebraic system $\mathbf{Y}\cong\mathbf{X}\beta$. Note that the t mesh was chosen so that every row of the matrix $\boldsymbol{X}$ subtends at least one quadrature panel. To obtain test problems in which the quadrature errors were completely negligible relative to the statistical errors, we did not use the $\boldsymbol{Y}$ vector computed from (2.3), but rather used $\boldsymbol{Y}^*$ defined by

$$\boldsymbol{Y}^* \equiv \boldsymbol{X}\boldsymbol{\beta}, \tag{2.6}$$

with the $\beta_j=\beta(\xi_j)$ calculated from (2.4). The calculation of $\boldsymbol{Y}^*$ used double-precision arithmetic, with the final results rounded to single precision on a machine with $\epsilon_{\text{mach}}\cong 7\times 10^{-15}$. The vector $\boldsymbol{\beta}$ can be regarded as the exact quantity to be estimated, and the vector $\boldsymbol{Y}^*$ can be regarded as the measurements that would be obtained if there were no measuring error.

To get the m random "measuring" errors ϵ_i, we let $\sigma_i=(10^{-6})Y_i^*$, $i=1,2,\ldots,m$, and picked sample vectors $\boldsymbol{\epsilon}$ from a multivariate normal distribution $N(\mathbf{0},\Sigma^2)$, with independently distributed elements—that is, $\Sigma^2=\operatorname{diag}\left(\sigma_1^2,\sigma_2^2,\ldots,\sigma_m^2\right)$. Adding the random errors chosen in this manner to the system (2.6) gave

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon}\sim N\left(\mathbf{0},\Sigma^2\right), \tag{2.7}$$

where $\boldsymbol{Y}$ is the vector of "measured" values. Because the calculation of the Y_i assures that (2.6) is satisfied to 14 digits, the standard deviations of $\boldsymbol{\epsilon}$ are all from 6 to 8 orders of magnitude greater than the corresponding truncation errors that arose in forming $\boldsymbol{Y}^*$. Thus the random errors were essentially the only errors in the problem.

2.1 PROBLEM I: AN OVERDETERMINED PROBLEM

For the first test problem we chose $m = 150$ and $n = 121$, producing a matrix $\boldsymbol{X}$ of rank n. For statistical studies, we generated several such problems by randomly choosing different error vectors ϵ, but picked a typical one for use in this article. The least squares solution

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\Sigma^{-2}\boldsymbol{X})^{-1}\boldsymbol{X}^T\Sigma^{-2}\boldsymbol{Y} \tag{2.8}$$

for this one is shown in the top of Figure 1, together with the exact solution $\beta(\xi)$. The maximum and minimum singular values of the scaled matrix $\Sigma^{-1}\boldsymbol{X}$ are $\sigma_1 \cong 3.3950 \times 10^9$ and $\sigma_{121} \cong 1.1610$ so the condition number is $\text{cond}(\Sigma^{-1}\boldsymbol{X}) \cong 2.924 \times 10^9$. Because $\epsilon_{\text{mach}} \cong 7 \times 10^{-15}$, we could reasonably expect to compute $\hat{\boldsymbol{\beta}}$ accurate to 6 digits. This means that all of the variation of the $\hat{\beta}(\xi_j)$ about the true solution $\beta(\xi)$ can be attributed to the measurement errors ϵ_i. This variation is surprisingly small in view of the large condition number of the matrix, but is still distressingly large when viewed as relative error in the solution. Many of the $\hat{\beta}_i$ are negative even though the exact solution $\beta(\xi)$ is everywhere nonnegative.

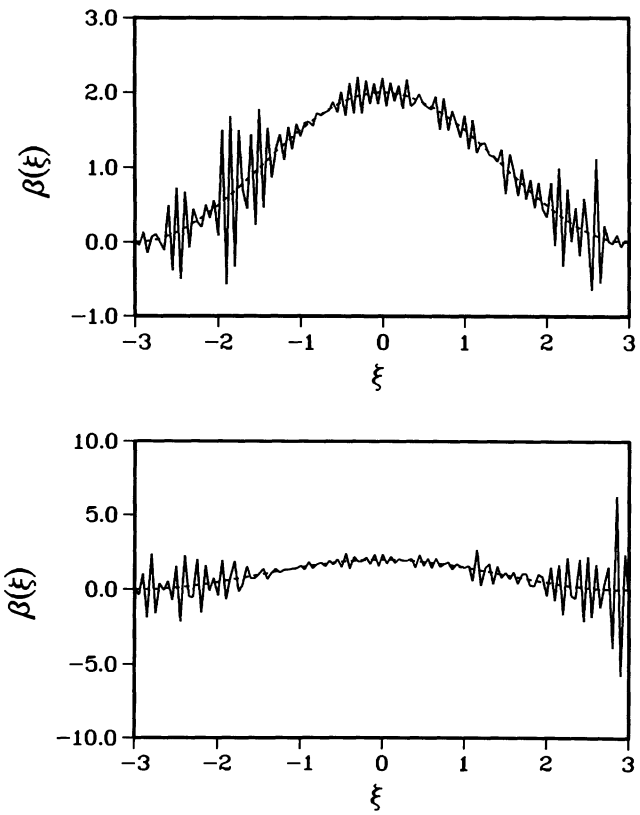

Figure 1. Classical Solutions to the Test Problems. The dashed lines are the exact solutions $\beta(\xi)$. Top: Least squares solution to Problem I. Bottom: Moore-Penrose generalized inverse solution to Problem II.

2.2 PROBLEM II: AN UNDERDETERMINED PROBLEM

Although the matrix in Problem I has a large condition number, the least squares problem is better than most that arise from first-kind integral equations. Many problems have rank deficient or even underdetermined matrices. If the kernel functions $X_i(\xi)$ are accessible, then it is always possible to choose $n < m$, but this may produce discretization errors larger than the measuring errors. Even if discretization error is acceptably small, choosing $n < m$ may still generate a rank-deficient problem. Many real-world problems are quite naturally underdetermined, with m and n being fixed by hardware considerations or other physical constraints. Often the analyst is given not the kernel functions, but rather the matrix $\boldsymbol{X}$ with $m < n$. To simulate such problems, we discretized the Phillips equation with $m = 108$ and $n = 121$. The matrix $\boldsymbol{X}$ has rank 108, and the maximum and minimum nonzero singular values of $\Sigma^{-1}\boldsymbol{X}$ are $\sigma_1 \cong 3.3728 \times 10^9$ and $\sigma_{108} \cong .14635$. The bottom frame of Figure 1 gives a plot of the generalized inverse solution

$$\hat{\boldsymbol{\beta}} = \left(\Sigma^{-1}\boldsymbol{X}\right)^{\dagger} \Sigma^{-1}\boldsymbol{Y}, \tag{2.9}$$

which oscillates about the true solution with even wider variations than those obtained for Problem I. This higher noise level is due mostly to the loss of information in using fewer observations. Of course the estimate is not unbiased, but the oscillations tend to be centered on the true curve, so the bias is evidently small relative to the random scatter.

2.3 WINDOW FUNCTIONS AND WINDOW VECTORS

The solutions presented in the preceding section correspond to the window functions

$$w_k(\xi) = \delta(\xi - \xi_k), \quad k = 1, 2, \ldots, n, \tag{2.10}$$

where $\delta(\xi - \xi_k)$ is the Dirac delta function centered on the quadrature mesh point ξ_k. The corresponding window responses are

$$\phi_k^* = \int_a^b \delta(\xi - \xi_k)\beta(\xi)\, d\xi = \beta_k, \tag{2.11}$$

and the corresponding window vectors are just the columns of the nth-order identity matrix. The estimates obtained from these windows were highly oscillatory because of the ill-conditioning of the regression model. These oscillations can be damped to some extent by seeking estimates for a set of average values of the function $\beta(\xi)$.

The simplest set of averages is defined by

$$\phi_k^* = \frac{1}{\Delta\xi} \int_{\xi_k}^{\xi_{k+1}} \beta(\xi)\, d\xi, \quad k = 1, 2, \ldots, n-1, \tag{2.12}$$

where $\Delta\xi = \xi_{k+1} - \xi_k$ is the quadrature mesh spacing. Each of these nonoverlapping averages is combined with its corresponding ξ interval to give a histogram approximation to $\beta(\xi)$. Using the trapezoidal rule to discretize the integrals, (2.12) gives $n - 1$ window vectors w_i, each with the ith and $(i+1)$st elements equal to $1/2$ and all other elements 0. We call this process *two-point averaging*.

Averaging can be extended to include more than one quadrature panel in each average, ending ultimately with *n-point averaging* that would seek to estimate only the average value of $\beta(\xi)$ on the whole interval $[a, b]$. In general, *ν-point averaging* uses window functions

$$\begin{aligned} w_k(\xi) &\equiv \tfrac{1}{(\nu-1)\Delta\xi},\ \xi_{(k-1)(\nu-1)+1} \le \xi \le \xi_{k(\nu-1)+1}, \quad k = 1, 2, \ldots, p \\ &\equiv 0, \quad \text{otherwise}, \qquad\qquad\qquad\qquad\quad k = 1, 2, \ldots, p, \end{aligned} \tag{2.13}$$

where p is the largest integer less than or equal to $(n-1)/(\nu-1)$. The typical window vector has the form

$$\mathbf{w}_k = \left(0, \ldots 0, \frac{1}{2(\nu-1)}, \frac{1}{\nu-1}, \frac{1}{\nu-1}, \ldots, \frac{1}{\nu-1}, \frac{1}{2(\nu-1)}, 0, \ldots, 0\right)^T, \tag{2.14}$$

with exactly ν nonzero elements beginning with element number $(k-1)(\nu-1)+1$. The corresponding histogram approximation will cover the whole interval $[a, b]$ only if $(n-1)$ is an exact multiple of $(\nu-1)$. Otherwise, there will be one or more (but fewer than ν) quadrature panels that will not be included in the set of averages.

3. CLASSICAL POINT ESTIMATION

To estimate the linear function

$$\phi^* = \boldsymbol{w}^T\boldsymbol{\beta} \tag{3.1}$$

subject to the statistical constraints

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathcal{E}(\boldsymbol{\epsilon}) = \mathbf{0}, \quad \mathcal{E}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right) = \Sigma^2, \tag{3.2}$$

we seek a linear estimator

$$\phi = \boldsymbol{u}^T\boldsymbol{Y}, \tag{3.3}$$

with mean and variance

$$\mathcal{E}(\phi) = \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{\beta}, \quad \text{var}(\phi) = \boldsymbol{u}^T\Sigma^2\boldsymbol{u}. \tag{3.4}$$

The estimator is unbiased iff

$$\boldsymbol{u}^T\boldsymbol{X} = \boldsymbol{w}^T, \tag{3.5}$$

which means the vector $\boldsymbol{w}$ is a linear combination of the rows of $\boldsymbol{X}$.

3.1 FULL RANK PROBLEMS: THE BEST LINEAR UNBIASED ESTIMATOR

If $\text{rank}(\boldsymbol{X}) = n \le m$, there will be many unbiased estimators

$$\boldsymbol{u}^T\boldsymbol{Y} = \left[\boldsymbol{w}^T\boldsymbol{X}^\dagger + \boldsymbol{z}^T\left(\boldsymbol{I}_m - \boldsymbol{X}\boldsymbol{X}^\dagger\right)\right]\boldsymbol{Y}, \tag{3.6}$$

where $X^\dagger$ is the generalized inverse of X and z is any m vector. The best linear unbiased estimator is given by the vector $\hat{u}$ that solves the constrained optimization problem

$$\operatorname{var}(\hat{u}^T Y) = \min_{u} \left\{ u^T \Sigma^2 u \mid u^T X = w^T \right\}. \tag{3.7}$$

It is easy to see that the solution is

$$\hat{u} = \Sigma^{-2} X (X^T \Sigma^{-2} X)^{-1} w, \quad \operatorname{var}(\hat{\phi}) = \hat{u}^T \Sigma^2 \hat{u} = w^T (X^T \Sigma^{-2} X)^{-1} w. \tag{3.8}$$

The best linear unbiased estimator can also be written

$$\hat{\phi} = w^T \hat{\beta}, \quad \hat{\beta} = (X^T \Sigma^{-2} X)^{-1} X^T \Sigma^{-2} Y, \tag{3.9}$$

where $\hat{\beta}$ is the solution vector for the weighted least squares problem

$$\hat{r}^2 = \min_{\underline{\beta}} \left\{ (Y - X\underline{\beta})^T \Sigma^{-2} (Y - X\underline{\beta}) \right\}. \tag{3.10}$$

As an example, consider Problem I (p. 71). If the set of window vectors is taken to be the columns of I_n, then the $\hat{\phi}_k$ are just the elements of the least squares solution $\hat{\beta}$ as in the top frame of Figure 1 (p. 71). The estimates corresponding to the two-point averaging window vectors are shown in the top frame of Figure 2. The noise level has been reduced by averaging, but this improvement is achieved only at the expense of a reduction in resolution for the independent variable ξ. Note also that the averaging does not prevent the occurrence of some negative estimates.

3.2 UNDERDETERMINED PROBLEMS

If X has less than full column rank, then the set of all least squares solutions can be written

$$\hat{\beta}(z) = \left(\Sigma^{-1} X\right)^\dagger \Sigma^{-1} Y + \left[I_n - \left(\Sigma^{-1} X\right)^\dagger \left(\Sigma^{-1} X\right) \right] z, \tag{3.11}$$

where z is an arbitrary n vector. An unbiased estimator for $w^T \beta$ exists only if w^T is an exact linear combination of the rows of X, in which case the best linear unbiased estimate is given by

$$\hat{u} = \Sigma^{-1} \left[\left(\Sigma^{-1} X\right)^\dagger \right]^T w, \quad \operatorname{var}(\hat{\phi}) = \hat{u}^T \Sigma^2 \hat{u} = w^T \left(X^T \Sigma^{-2} X \right)^\dagger w, \tag{3.12}$$

so

$$\hat{\phi} = w^T \left(\Sigma^{-1} X\right)^\dagger \Sigma^{-1} Y = w^T \hat{\beta}, \tag{3.13}$$

where $\hat{\beta}$ is the minimal length solution (2.9). Being limited to window vectors expressible as linear combinations of the rows of X places severe restrictions on the information that can be elicited about the unknown function $\beta(\xi)$. This restriction is imposed in the method of Backus and Gilbert that constructs just such a set of window vectors, called *averaging kernels* (Tarantola 1987). Ideally, each averaging kernel should resemble a

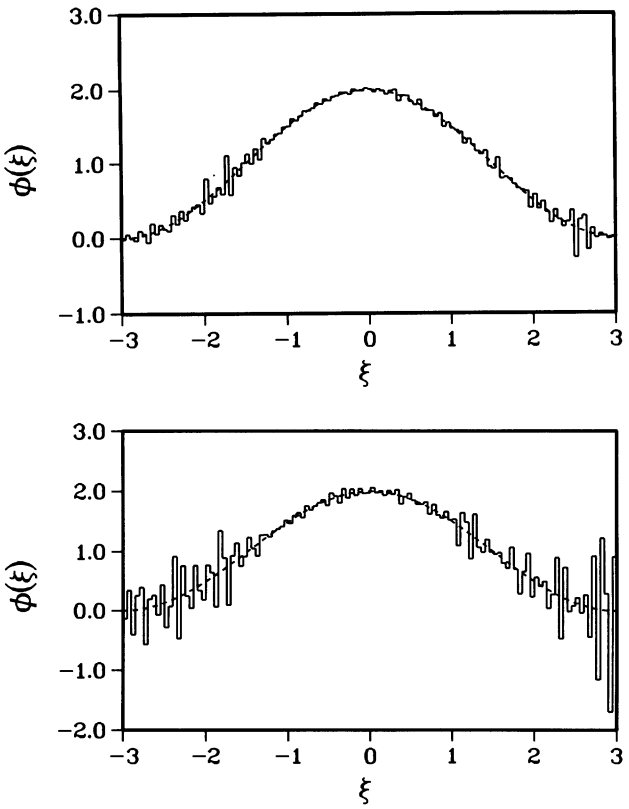

Figure 2. Point Estimates Using Two-point Averaging. The dashed lines are the exact solutions $\beta(\xi)$. The solid lines are histogram of the estimates $\hat{\phi}_k$ plotted over the corresponding intervals $[\xi_k, \xi_{k+1}]$. Top: Best linear unbiased estimates for Problem I. Bottom: Moore-Penrose generalized inverse estimates for Problem II.

narrow Gaussian centered on the point ξ_k. In practice the rows of $\boldsymbol{X}$ rarely admit good resemblances to Gaussians, so the $\boldsymbol{w}_k$ are usually nonsymmetric weighting functions that change from one ξ_k to the next. The resulting set of point estimates $(\xi_k, \hat{\phi}_k)$ comprise a smoothed discrete approximation to the unknown function $\beta(\xi)$. This smoothing is difficult to characterize and nonuniform over the range of ξ.

If $\boldsymbol{w}$ is not expressible as a linear combination of rows of $\boldsymbol{X}$, then the average $\boldsymbol{w}^T\boldsymbol{\beta}$ is said to be *inestimable*. This does not mean that no unbiased estimate exists, but rather that there exists no unbiased estimator of the form $\boldsymbol{u}^T\boldsymbol{Y}$. In theory, there exist unbiased estimates of the form $\boldsymbol{w}^T\boldsymbol{\beta}(\tilde{z})$, where $\boldsymbol{\beta}(\tilde{z})$ is some solution (3.11) of the underdetermined least squares problem. In reality, it is necessary to accept the fact that computable estimates will be biased and then try to assure that the errors due to bias are small in comparison to the random variations introduced by the $\boldsymbol{\epsilon}$.

In Section 2 we presented the generalized inverse solution for Problem II. It is apparent from the bottom frame of Figure 1 (p. 71) that any bias in this approximation is smaller than the scatter due to the random errors. In the bottom frame of Figure 2 we give the histogram approximation corresponding to the two-point averaging windows. Even though the averaging considerably reduces the scatter due to the random errors, it

is still not possible to detect the bias. It would seem that the bias is not so large that it rules out the possibility of obtaining useful information about the function $\beta(\xi)$.

4. CLASSICAL ONE-AT-A-TIME CONFIDENCE INTERVALS

For any unbiased estimate, one-at-a-time confidence intervals for ϕ^* can be constructed from the transformed random variable

$$\eta = \frac{\phi - \phi^*}{\sqrt{\text{var}(\phi)}} = \frac{\boldsymbol{u}^T\boldsymbol{Y} - \boldsymbol{w}^T\boldsymbol{\beta}}{\sqrt{\boldsymbol{u}^T\Sigma^2\boldsymbol{u}}}, \tag{4.1}$$

which has mean $\mathcal{E}(\eta) = 0$ and variance $\text{var}(\eta) = 1$. For a given probability α, a $100\alpha\%$ confidence interval for ϕ^* can be constructed by determining a number κ such that $\Pr\{-\kappa \le \eta \le +\kappa\} = \alpha$. Substituting (4.1) and rearranging gives

$$\Pr\left\{\left(\boldsymbol{u}^T\boldsymbol{Y} - \kappa\sqrt{\boldsymbol{u}^T\Sigma^2\boldsymbol{u}}\right) \le \boldsymbol{w}^T\boldsymbol{\beta} \le \left(\boldsymbol{u}^T\boldsymbol{Y} + \kappa\sqrt{\boldsymbol{u}^T\Sigma^2\boldsymbol{u}}\right)\right\} = \alpha, \tag{4.2}$$

so the interval

$$[\phi^{\text{lo}}, \phi^{\text{up}}] = \left[\left(\boldsymbol{u}^T\boldsymbol{Y} - \kappa\sqrt{\boldsymbol{u}^T\Sigma^2\boldsymbol{u}}\right), \left(\boldsymbol{u}^T\boldsymbol{Y} + \kappa\sqrt{\boldsymbol{u}^T\Sigma^2\boldsymbol{u}}\right)\right] \tag{4.3}$$

is a $100\alpha\%$ confidence interval for ϕ^*.

Given the value κ, the foregoing results are valid for any unbiased estimator $\boldsymbol{u}^T\boldsymbol{Y}$. The best linear unbiased estimator gives the bounds

$$\begin{aligned} \hat{\phi}^{\text{lo}} &= \hat{\boldsymbol{u}}^T\boldsymbol{Y} - \kappa\sqrt{\hat{\boldsymbol{u}}^T\Sigma^2\hat{\boldsymbol{u}}} = \boldsymbol{w}^T\hat{\boldsymbol{\beta}} - \kappa\sqrt{\boldsymbol{w}^T(\boldsymbol{X}^T\Sigma^{-2}\boldsymbol{X})^{-1}\boldsymbol{w}} \\ \hat{\phi}^{\text{up}} &= \hat{\boldsymbol{u}}^T\boldsymbol{Y} + \kappa\sqrt{\hat{\boldsymbol{u}}^T\Sigma^2\hat{\boldsymbol{u}}} = \boldsymbol{w}^T\hat{\boldsymbol{\beta}} + \kappa\sqrt{\boldsymbol{w}^T(\boldsymbol{X}^T\Sigma^{-2}\boldsymbol{X})^{-1}\boldsymbol{w}}, \end{aligned} \tag{4.4}$$

where $\hat{\boldsymbol{\beta}}$ is the least squares solution vector. The width of the confidence interval is directly proportional to the value of κ. Optimally narrow confidence intervals can be computed only if the probability density function for η is known, and this requires a knowledge of the joint probability density function for $\boldsymbol{\epsilon}$.

4.1 CONFIDENCE INTERVALS FOR NORMALLY DISTRIBUTED ERRORS

In many applications, $\boldsymbol{\epsilon}$ is known (or assumed) to have a multivariate normal distribution, $\boldsymbol{\epsilon} \sim \text{N}(\boldsymbol{0}, \Sigma^2)$, so η follows the standard normal distribution $\eta \sim n(0,1)$. This means that for any α $(0 < \alpha < 1)$, it is possible to find a corresponding value κ $(0 < \kappa < \infty)$ satisfying

$$\frac{1}{\sqrt{2\pi}}\int_{-\kappa}^{+\kappa} \exp\left(-\frac{\eta^2}{2}\right) d\eta = \alpha. \tag{4.5}$$

As an example, consider Problem I (p. 71) and the best linear unbiased estimators for the two-point averaging window vectors. Using $\kappa = 1.960$ for 95% intervals, Equation (4.4) defines the confidence bounds plotted as solid lines in the top frame of Figure 3.

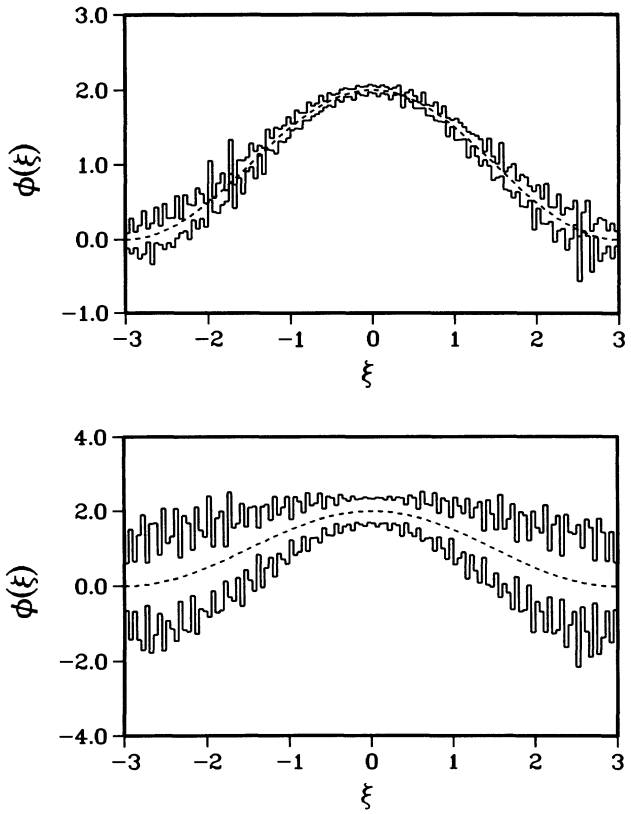

Figure 3. Classical 95% Confidence Intervals for Problem I. The dashed lines are the exact solution $\beta(\xi)$. The solid lines are histograms of lower and upper bounds for two-point averaging windows. Top: One-at-a-time confidence intervals. Bottom: Simultaneous confidence intervals.

4.2 CONFIDENCE INTERVALS FROM CHEBYSHEV'S INEQUALITY

If the joint probability density function for ϵ is not known, then valid, suboptimal confidence intervals can be constructed by using Chebyshev's inequality (Hogg and Craig 1965, chap. 1), which guarantees that, for any $\kappa > 0$,

$$\Pr\{|\ \eta - \mathcal{E}(\eta)\ | \geq\ \kappa \mathrm{var}(\eta)\} \leq \frac{1}{\kappa^2}. \tag{4.6}$$

It follows then that, for any unbiased estimator $\boldsymbol{u}^T\boldsymbol{Y}$,

$$\Pr\left\{|\ \boldsymbol{u}^T\boldsymbol{Y} - \boldsymbol{w}^T\boldsymbol{\beta}\ | < \kappa\sqrt{\boldsymbol{u}^T\Sigma^2\boldsymbol{u}}\right\} \geq 1 - \frac{1}{\kappa^2}. \tag{4.7}$$

Therefore, for any α $(0 < \alpha < 1)$, if $\kappa = 1/\sqrt{1-\alpha}$, then

$$\Pr\left\{\left(\boldsymbol{u}^T\boldsymbol{Y} - \kappa\sqrt{\boldsymbol{u}^T\Sigma^2\boldsymbol{u}}\right) < \boldsymbol{w}^T\boldsymbol{\beta} < \left(\boldsymbol{u}^T\boldsymbol{Y} + \kappa\sqrt{\boldsymbol{u}^T\Sigma^2\boldsymbol{u}}\right)\right\} \geq \alpha, \tag{4.8}$$

so the interval (4.3) is a $100\alpha\%$ confidence interval for ϕ^*.

Confidence intervals calculated from Chebyshev's inequality are very conservative because they must be wide enough to accommodate any possible probability density for

Table 1. κ Values for Chebyshev Inequality and Standard Normal Distribution

α	κ *Chebyshev*	κ *normal*	*ratio*
.6667	1.732	.967	1.79
.95	4.472	1.960	2.28
.99	10.000	2.575	3.88
.999	31.622	3.295	9.60

η. Table 1 compares the sizes of Chebyshev intervals and normal distribution intervals for some commonly used confidence levels. The ratio is the factor by which the interval width must be expanded if the form of the error distribution is not known.

5. THE GEOMETRY OF CONFIDENCE INTERVAL ESTIMATION

5.1 FULL RANK PROBLEMS

No matter what the relation between α and κ, the corresponding confidence interval bounds for the best linear unbiased estimator can be calculated from (4.4). It is not difficult to show that the endpoints of this interval are also defined by the two constrained estimation problems

$$\begin{aligned} \hat{\phi}^{\text{lo}} &= \min_{\underline{\beta}} \left\{ \boldsymbol{w}^T \underline{\beta} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\beta})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\beta}) = \hat{r}^2 + \kappa^2 \right\} \\ \hat{\phi}^{\text{up}} &= \max_{\underline{\beta}} \left\{ \boldsymbol{w}^T \underline{\beta} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\beta})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\beta}) = \hat{r}^2 + \kappa^2 \right\}, \end{aligned} \tag{5.1}$$

where $\hat{r}^2$ is the minimum sum of squares (3.10). These problems can be solved by Lagrange multipliers to give solution vectors

$$\hat{\boldsymbol{\beta}}^{\text{lo}} = \hat{\boldsymbol{\beta}} - \frac{\kappa}{\sqrt{\boldsymbol{w}^T \left(\boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X} \right)^{-1} \boldsymbol{w}}} \left(\boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X} \right)^{-1} \boldsymbol{w} \tag{5.2}$$

and

$$\hat{\boldsymbol{\beta}}^{\text{up}} = \hat{\boldsymbol{\beta}} + \frac{\kappa}{\sqrt{\boldsymbol{w}^T \left(\boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X} \right)^{-1} \boldsymbol{w}}} \left(\boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X} \right)^{-1} \boldsymbol{w}, \tag{5.3}$$

and values for $\hat{\phi}^{\text{lo}}$ and $\hat{\phi}^{\text{up}}$ that are the same as the values given by (4.4).

The common constraint region for problems (5.1) can be written

$$\begin{aligned} \mathcal{S}(\kappa) &= \left\{ \underline{\beta} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\beta})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\beta}) = \hat{r}^2 + \kappa^2 \right\} \\ &= \left\{ \underline{\beta} \mid (\underline{\beta} - \hat{\beta})^T \boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X} (\underline{\beta} - \hat{\beta}) = \kappa^2 \right\}, \end{aligned} \tag{5.4}$$

which defines the surface of an ellipsoid centered on the least squares solution $\hat{\beta}$ (Rust and Burrus 1972, chap. 2). The size of this ellipsoid varies with κ. For any value of the

parameter ϕ, the set of points $\underline{\beta}$ satisfying $\boldsymbol{w}^T\underline{\boldsymbol{\beta}} = \phi$ forms a hyperplane orthogonal to the vector $\boldsymbol{w}$. Therefore, the bounds $\hat{\phi}^{\text{lo}}$ and $\hat{\phi}^{\text{up}}$ are the values of ϕ on the two tangential support planes of the ellipsoid $\mathcal{S}(\kappa)$ that are orthogonal to $\boldsymbol{w}$.

The location of the center of the ellipsoid $\mathcal{S}(\kappa)$ depends on $\boldsymbol{Y}$, and its size is scaled by the value of κ, but its orientation and shape are completely determined by the matrix $\boldsymbol{X}^T\Sigma^{-2}\boldsymbol{X}$. Let the singular value decomposition of $\Sigma^{-1}\boldsymbol{X}$ be

$$\Sigma^{-1}\boldsymbol{X} = \boldsymbol{U}\begin{pmatrix} \Lambda \\ \boldsymbol{O} \end{pmatrix}\boldsymbol{V}^T, \tag{5.5}$$

with

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_m = \boldsymbol{U}\boldsymbol{U}^T, \quad \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_n = \boldsymbol{V}\boldsymbol{V}^T, \quad \Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), \tag{5.6}$$

with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. It follows then that

$$\boldsymbol{X}^T\Sigma^{-2}\boldsymbol{X} = \boldsymbol{V}\Lambda^2\boldsymbol{V}^T. \tag{5.7}$$

It is easy to see that the mutually orthogonal vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ define the directions of the major axes of $\mathcal{S}(\kappa)$, and the lengths of those axes are given by

$$\ell_i = \frac{2\kappa}{\lambda_i}, \quad i = 1, 2, \ldots, n. \tag{5.8}$$

From (4.4), (5.7), and the definition of Λ, it follows that for any window vector $\boldsymbol{w}$, the corresponding α-level confidence interval has width

$$\hat{\phi}^{\text{up}} - \hat{\phi}^{\text{lo}} = 2\kappa\left[\sum_{j=1}^{n}\frac{1}{\lambda_j^2}\left(\boldsymbol{v}_j^T\boldsymbol{w}\right)^2\right]^{\frac{1}{2}} = \left[\sum_{j=1}^{n}\ell_i^2\left(\boldsymbol{w}^T\boldsymbol{v}_j\right)^2\right]^{\frac{1}{2}}. \tag{5.9}$$

Because $\boldsymbol{v}_j$ is a unit vector (i.e., $\boldsymbol{v}_j^T\boldsymbol{v}_j = 1$), the scalar $\boldsymbol{w}^T\boldsymbol{v}_j$ is just the projection of $\boldsymbol{w}$ on the jth major axis of the ellipsoid, and ℓ_j is the length of that major axis.

Linear regression models obtained from first-kind integral equations are almost always poorly conditioned. The major axes of $\mathcal{S}(\kappa)$ corresponding to the smaller singular values are, therefore, usually greatly elongated. From (5.9) it follows that if $\boldsymbol{w}$ has a nonzero projection on any of the singular vectors corresponding to the smaller singular values, then the confidence interval for $\boldsymbol{w}^T\boldsymbol{\beta}$ will be very wide. Almost every window vector designed to elicit information about the function $\beta(\xi)$ will have nonzero components in the directions of these elongated axes. In the extreme case, one or more of the singular values are 0 and the corresponding axes of $\mathcal{S}(\kappa)$ are infinite in length, yielding infinite intervals. When such a problem is read into a computer, the machine truncation errors almost always produce a full rank matrix on the machine, but there is no danger of obtaining deceptively good confidence intervals as a result. If ϵ_M is the machine epsilon, then typically for a ϕ_k^* of order unity the classical estimates (4.4) give intervals like $[-\mathcal{O}(\epsilon_M^{-1}), +\mathcal{O}(\epsilon_M^{-1})]$ which, though shorter than $[-\infty, +\infty]$, are nevertheless useless.

5.2 UNDERDETERMINED PROBLEMS

For problems with $m < n$, the matrix is exactly rank deficient with $\lambda_{m+1} = \lambda_{m+2} = \cdots = \lambda_n = 0$. The ellipsoid $\mathcal{S}(\kappa)$ has at least $n - m$ infinitely long major axes, and its center is not a single point $\hat{\boldsymbol{\beta}}$ but rather a coset of the subspace spanned by the singular vectors $\boldsymbol{v}_{m+1}, \boldsymbol{v}_{m+2}, \ldots, \boldsymbol{v}_n$—more precisely, the set of points defined by (3.11). In that expression, the set of points

$$\mathcal{N}(\Sigma^{-1}\boldsymbol{X}) = \left\{ \left[\boldsymbol{I}_n - \left(\Sigma^{-1}\boldsymbol{X}\right)^{\dagger} \left(\Sigma^{-1}\boldsymbol{X}\right) \right] \boldsymbol{z} \mid \boldsymbol{z} \quad \text{arbitrary} \right\} \tag{5.10}$$

is the subspace, and

$$\hat{\boldsymbol{\beta}} = \left(\Sigma^{-1}\boldsymbol{X}\right)^{\dagger} \Sigma^{-1}\boldsymbol{Y} \tag{5.11}$$

is the displacement of the coset from the origin. If $\boldsymbol{w}$ is orthogonal to the subspace $\mathcal{N}(\Sigma^{-1}\boldsymbol{X})$, then it is an exact linear combination of the rows of $\boldsymbol{X}$, so the best linear unbiased estimators for $\boldsymbol{w}^T\boldsymbol{\beta}$ and its variance are given by (3.12) and (3.13). The confidence interval bounds are

$$\begin{aligned} \hat{\phi}^{\text{lo}} &= \boldsymbol{w}^T\hat{\boldsymbol{\beta}} - \kappa\sqrt{\boldsymbol{w}^T \left(\boldsymbol{X}^T\Sigma^{-2}\boldsymbol{X}\right)^{\dagger} \boldsymbol{w}} \\ \hat{\phi}^{\text{up}} &= \boldsymbol{w}^T\hat{\boldsymbol{\beta}} + \kappa\sqrt{\boldsymbol{w}^T \left(\boldsymbol{X}^T\Sigma^{-2}\boldsymbol{X}\right)^{\dagger} \boldsymbol{w}}, \end{aligned} \tag{5.12}$$

where $\hat{\boldsymbol{\beta}}$ is the minimal length solution (5.11).

If $\boldsymbol{w}$ has a nonzero projection on any of the infinite axes of $\mathcal{S}(\kappa)$, then $\boldsymbol{w}^T\boldsymbol{\beta}$ is said to be inestimable because there exists no unbiased estimator of the form $\phi = \boldsymbol{u}^T\boldsymbol{Y}$. The center coset does contain at least one point

$$\underline{\boldsymbol{\beta}}(\tilde{\boldsymbol{z}}) = \left(\Sigma^{-1}\boldsymbol{X}\right)^{\dagger} \Sigma^{-1}\boldsymbol{Y} + \left[\boldsymbol{I}_n - \left(\Sigma^{-1}\boldsymbol{X}\right)^{\dagger} \left(\Sigma^{-1}\boldsymbol{X}\right) \right] \tilde{\boldsymbol{z}}, \tag{5.13}$$

for which $\boldsymbol{w}^T\underline{\boldsymbol{\beta}}(\tilde{\boldsymbol{z}})$ gives an unbiased estimate. Taking $\tilde{\boldsymbol{z}} = \boldsymbol{\beta}$ would give such an estimate, but of course $\boldsymbol{\beta}$ is unknown. Even if it were possible to find some $\tilde{\boldsymbol{z}}$ that gives an unbiased estimate, the confidence interval for $\boldsymbol{w}^T\boldsymbol{\beta}$ would still be unbounded.

6. CLASSICAL SIMULTANEOUS CONFIDENCE INTERVALS

The method that we shall use for estimating a set of α-level simultaneous confidence intervals for a whole set of window vectors is based on the α-level confidence ellipsoid for the unknown vector $\boldsymbol{\beta}$. This ellipsoid is centered at the least squares solution $\hat{\boldsymbol{\beta}}$, with size determined by a parameter γ_n^2 chosen so that

$$\Pr\left\{ \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T\Sigma^{-2}\boldsymbol{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) \le \gamma_n^2 \right\} = \alpha. \tag{6.1}$$

The required set of simultaneous confidence intervals is constructed from the support

planes for this ellipsoid. In particular, the set of intervals defined by

$$\begin{aligned}\hat{\phi}_k^{\text{lo}} &= \min_{\underline{\boldsymbol{\beta}}} \left\{ \boldsymbol{w}_k^T \underline{\boldsymbol{\beta}} \mid \left(\underline{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X} \left(\underline{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right) = \gamma_n^2 \right\}, \quad k = 1, 2, \ldots, p \\ \hat{\phi}_k^{\text{up}} &= \max_{\underline{\boldsymbol{\beta}}} \left\{ \boldsymbol{w}_k^T \underline{\boldsymbol{\beta}} \mid \left(\underline{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X} \left(\underline{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right) = \gamma_n^2 \right\}, \quad k = 1, 2, \ldots, p, \end{aligned} \tag{6.2}$$

satisfy

$$\Pr\left\{ \hat{\phi}_k^{\text{lo}} \leq \phi_k^* \leq \hat{\phi}_k^{\text{up}}, \quad k = 1, 2, \ldots, p \right\} \geq \alpha, \tag{6.3}$$

and this result holds for any number of averaging vectors. These simultaneous confidence intervals are defined by the support planes of an ellipsoid in the same way as are the one-at-a-time intervals given by (5.1). The only difference is in the size of the defining ellipsoid. It is easy to see that if $\text{rank}(\boldsymbol{X}) = n \leq m$, then the solutions to the constrained optimization problems (6.2) are

$$\begin{aligned}\hat{\phi}_k^{\text{lo}} &= \boldsymbol{w}_k^T \hat{\boldsymbol{\beta}} - \gamma_n \sqrt{\boldsymbol{w}_k^T \left(\boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X}\right)^{-1} \boldsymbol{w}_k}, \quad k = 1, 2, \ldots, p \\ \hat{\phi}_k^{\text{up}} &= \boldsymbol{w}_k^T \hat{\boldsymbol{\beta}} + \gamma_n \sqrt{\boldsymbol{w}_k^T \left(\boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X}\right)^{-1} \boldsymbol{w}_k}, \quad k = 1, 2, \ldots, p, \end{aligned} \tag{6.4}$$

where $\hat{\boldsymbol{\beta}}$ is the least squares solution vector (3.9).

An alternative to basing simultaneous confidence intervals on the underlying confidence ellipsoid is to use Bonferroni-type inequalities to combine a collection of one-at-a-time intervals into a set of simultaneous ones (Tong 1980). See, for example, the work of Hunter (1976) that is discussed by Stoline (1983), the work of Worsley (1982), and the methods devised by Slepian and Šidák and discussed by Tong (1980, chap. 2). These methods construct a simultaneous set with a given probability by requiring an appropriately higher probability for each of the one-at-a-time intervals gathered into the set. We will not develop inequality-constrained Bonferroni methods in this article, but note that the results we shall obtain in Section 8 will be useful in any development of nonnegatively constrained Bonferroni intervals.

6.1 SIMULTANEOUS INTERVALS FOR NORMALLY DISTRIBUTED ERRORS

If $\boldsymbol{\epsilon}$ is normally distributed and $\boldsymbol{X}$ has full column rank, then

$$\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \chi^2(n),$$

so

$$\Pr\left\{ \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \leq \gamma_n^2 \right\} = \int_0^{\gamma_n^2} \chi^2(n; q) dq \equiv \alpha, \tag{6.5}$$

where $\chi^2(n; q)$ is the probability density function for the $\chi^2(n)$ distribution. It is interesting to compare the widths of the simultaneous intervals (6.4) with those of the

Table 2. The Ratio γ_n/κ for Normally Distributed Errors

α	$n = 20$	$n = 30$	$n = 50$	$n = 70$	$n = 100$	$n = 150$	$n = 200$
.6667	4.9	5.9	7.6	8.9	10.6	13.0	14.9
.95	2.9	3.4	4.2	4.9	5.7	6.8	7.8
.99	2.4	2.8	3.4	3.9	4.5	5.4	6.1
.999	2.0	2.3	2.8	3.2	3.7	4.4	4.9

one-at-a-time intervals (4.4). The ratio of the interval widths is tabulated in Table 2 for several values of α and n.

As an example, we again consider Problem I with two-point averaging windows and seek 95% simultaneous intervals: $\gamma_n^2 = 12.14$. Using this value in (6.4) gives the intervals shown in the bottom frame of Figure 3 (p. 77). Comparing this plot with the one in the top frame gives a rough idea of the price that must be paid for the additional certainty of simultaneous intervals.

6.2 SIMULTANEOUS INTERVALS FROM CHEBYSHEV'S INEQUALITY

If the form of the joint probability density function for ϵ is unknown, then valid, though conservative, confidence ellipsoids can be constructed from Chebyshev's inequality. For a given α, it is necessary to determine a corresponding γ_n such that

$$\Pr\{\boldsymbol{\beta} \in \mathcal{S}(\gamma_n)\} = \Pr\left\{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \leq \gamma_n^2\right\} \geq \alpha. \tag{6.6}$$

A multivariate Chebyshev inequality given by Olkin and Pratt (1958) ensures that for any α, $(0 < \alpha < 1)$, this will be satisfied if $\gamma_n = n/\sqrt{1-\alpha}$. In general, however, the demand for the certainty of simultaneous intervals coupled with a lack of knowledge about the distribution of the errors usually produces intervals so wide that they are almost useless.

7. NONNEGATIVELY CONSTRAINED ESTIMATION

The least squares problem arising from the discretization of a system of first-kind integral equations is usually poorly conditioned with respect to small variations in the measured vector $\boldsymbol{Y}$. Quite often, the exact matrix $\boldsymbol{X}$ is rank deficient which means, in the classical theory, that the linear functions $\phi_k^* = \boldsymbol{w}_k^T \boldsymbol{\beta}$ may not even be estimable. Fortunately, in cases where the solution $\beta(\xi)$ of the integral equations is physically constrained to be nonnegative, the imposition of those constraints on the solution vector $\hat{\boldsymbol{\beta}}$ almost always stabilizes the estimation problems and gives biased, but bounded, physically plausible solutions, no matter what the value of rank($\boldsymbol{X}$). Therefore we append the constraint $\boldsymbol{\beta} \geq \boldsymbol{0}$ to the linear estimation model (1.2)—that is,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathcal{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}, \quad \mathcal{E}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right) = \Sigma^2, \quad \Pr\{\boldsymbol{\beta} \geq \boldsymbol{0}\} = 1.0. \tag{7.1}$$

7.1 CONSTRAINED POINT ESTIMATION

When the nonnegativity constraint is appended, the point-estimation problem becomes

$$\begin{aligned}\rho_{\min} &= \min_{\underline{\beta}\geq\mathbf{0}} \left\{(\boldsymbol{X}\underline{\beta}-\boldsymbol{Y})^T\Sigma^{-2}(\boldsymbol{X}\underline{\beta}-\boldsymbol{Y})\right\} \\ &= \hat{r}^2 + \min_{\underline{\beta}\geq\mathbf{0}} \left\{(\underline{\beta}-\hat{\beta})^T\boldsymbol{X}^T\Sigma^{-2}\boldsymbol{X}(\underline{\beta}-\hat{\beta})\right\}, \end{aligned} \tag{7.2}$$

with $\hat{r}^2$ defined by (3.10). If $\check{\beta}$ denotes the solution vector—that is, if

$$\rho_{\min} = (\boldsymbol{X}\check{\beta}-\boldsymbol{Y})^T\Sigma^{-2}(\boldsymbol{X}\check{\beta}-\boldsymbol{Y}), \tag{7.3}$$

then estimates of the ensemble of averages can be computed by

$$\check{\phi}_k = \boldsymbol{w}_k^T\check{\beta}, \quad k = 1,2,\ldots,p. \tag{7.4}$$

Equation (7.2) defines a nonnegatively constrained least squares problem for which there is no explicit closed form solution, but good algorithms for the numerical solution

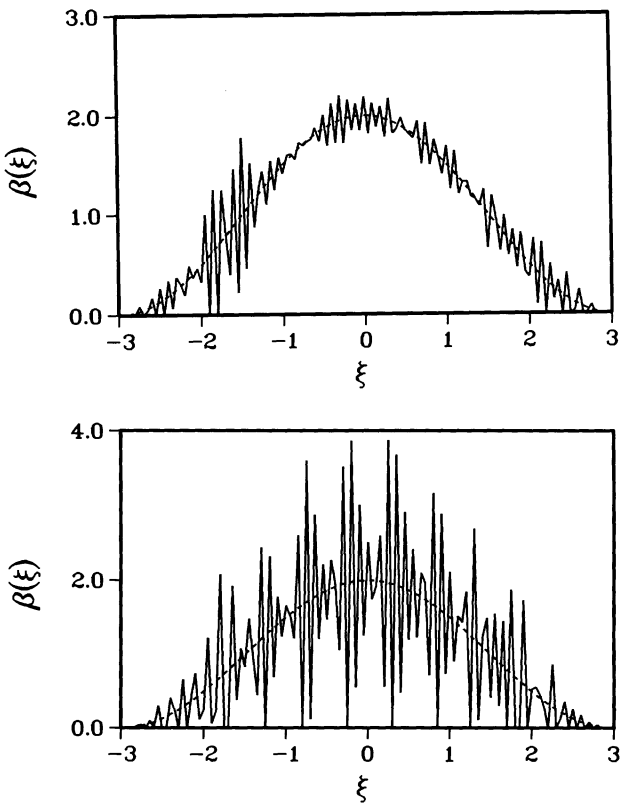

Figure 4. Nonnegatively Constrained Solutions to the Test Problems. The dashed lines are the exact solutions $\beta(\xi)$, and the jagged solid lines are plots of $\check{\beta}_k$. Top: Problem I. Bottom: Problem II.

have been available since the mid-1970s (Lawson and Hanson 1974). Figure 4 gives plots of these solutions for Problems I and II. Comparison of these plots with those in Figure 1 (p. 71) shows that the nonnegativity constraints gave good, but not spectacular, improvement. For most real-world ill-posed problems, the improvement is much more dramatic.

7.2 CONSTRAINED INTERVAL ESTIMATION

The classical method for estimating confidence intervals can also be extended to take the nonnegativity constraints into account. For both one-at-a-time and simultaneous intervals, the classical, unconstrained interval estimation problems can be written in the form

$$\begin{aligned} \hat{\phi}^{\text{lo}} &= \min_{\underline{\boldsymbol{\beta}}} \left\{ \boldsymbol{w}^T \underline{\boldsymbol{\beta}} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) = \omega^2 \right\} \\ \hat{\phi}^{\text{up}} &= \max_{\underline{\boldsymbol{\beta}}} \left\{ \boldsymbol{w}^T \underline{\boldsymbol{\beta}} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) = \omega^2 \right\}, \end{aligned} \tag{7.5}$$

where ω^2 is a constant chosen to give the desired confidence level α. We will show in the following that valid nonnegatively constrained confidence intervals can be obtained by solving problems of the form

$$\begin{aligned} \check{\phi}^{\text{lo}} &= \min_{\underline{\boldsymbol{\beta}}} \left\{ \boldsymbol{w}^T \underline{\boldsymbol{\beta}} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) = \mu^2, \quad \underline{\boldsymbol{\beta}} \geq \boldsymbol{0} \right\} \\ \check{\phi}^{\text{up}} &= \max_{\underline{\boldsymbol{\beta}}} \left\{ \boldsymbol{w}^T \underline{\boldsymbol{\beta}} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) = \mu^2, \quad \underline{\boldsymbol{\beta}} \geq \boldsymbol{0} \right\}, \end{aligned} \tag{7.6}$$

where $\mu = \omega$ for simultaneous confidence intervals, but μ differs from ω for one-at-a-time intervals.

The common constraint region in (7.6) is the intersection of the n-dimensional μ ellipsoid with the positive orthant. For any reasonable discretization of the integral equations, the value of n will be large enough to assure that the positive orthant constitutes a very small fraction of n space. The resulting intersection is generally much smaller in all directions than is the μ ellipsoid itself. Even for rank-deficient matrices $\boldsymbol{X}$ this intersection is almost always bounded. If $\boldsymbol{X}$ is rank deficient, then one or more of the smallest eigenvalues of $\boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X}$ will be 0, and the ellipsoid will be unbounded in the directions of the corresponding eigenvectors. The intersection of the ellipsoid with the positive orthant will, however, be unbounded only if at least one of these degenerate eigenvectors lies in the positive orthant. Because $\underline{\boldsymbol{\beta}}$ space has 2^n orthants and $\boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X}$ has fewer than n degenerate eigenvalues, it will be a very rare occurrence for one of the corresponding eigenvectors to extend into the positive orthant. In many physical applications the matrix $\boldsymbol{X}$ is nonnegative, and in such cases $\underline{\boldsymbol{\beta}} \geq \boldsymbol{0}$, $\underline{\boldsymbol{\beta}} \neq \boldsymbol{0} \Rightarrow \boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X} \underline{\boldsymbol{\beta}} \neq 0 \cdot \underline{\boldsymbol{\beta}}$; thus, for these applications, no degenerate eigenvector could lie in the positive orthant.

It is not possible to write closed-form solutions for the problems (7.6), but numerical calculation of the solutions is possible. The first person to employ nonnegativity constraints to reduce the size of estimated confidence intervals was Walter R. Burrus

(1965) who used the technique in neutron and gamma-ray spectrum unfolding problems. Burrus did not give an algorithm for solving (7.6), but he and his colleagues (Burrus, Rust, and Cope 1980; Cope and Rust 1979; Pierce and Rust 1985; Rust and Burrus 1971, 1972) developed several algorithms for computing suboptimal interval approximations to $[\check{\phi}^{\text{lo}}, \check{\phi}^{\text{up}}]$. In spite of the suboptimality, the intervals thus obtained were uniformly much better than the classical intervals. O'Leary and Rust (1986) developed an algorithm called BRACKET-LS for computing the optimal intervals. This algorithm was used to compute the confidence intervals given in the remainder of this article.

7.3 SIMULTANEOUS CONFIDENCE INTERVALS

The extension of the classical theory to accommodate nonnegativity constraints in the estimation of one-at-a-time intervals is complicated, so we defer it to the next section. The extension to estimating simultaneous intervals is straightforward because the intersection of a confidence ellipsoid with the positive orthant is itself a confidence region with the same confidence level as the ellipsoid. Let γ be chosen so that

$$\begin{aligned} \mathcal{S}(\gamma) &= \left\{\underline{\boldsymbol{\beta}} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) \le \hat{r}^2 + \gamma^2\right\} \\ &= \left\{\underline{\boldsymbol{\beta}} \mid \left(\underline{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X} \left(\underline{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right) \le \gamma^2\right\} \end{aligned} \tag{7.7}$$

is a $100\alpha\%$ confidence ellipsoid for $\boldsymbol{\beta}$, and define events $\mathcal{A}$ and $\mathcal{B}$ by

$$\mathcal{A} = \left\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \le \hat{r}^2 + \gamma^2\right\}, \quad \mathcal{B} = \{\boldsymbol{\beta} \ge \boldsymbol{0}\}, \tag{7.8}$$

with probabilities $\mathrm{P}(\mathcal{A}) = \alpha$ and $\mathrm{P}(\mathcal{B}) = 1.0$. It follows from $\mathrm{P}(\mathcal{A} \cap \mathcal{B}) = \mathrm{P}(\mathcal{A}) + \mathrm{P}(\mathcal{B}) - \mathrm{P}(\mathcal{A} \cup \mathcal{B})$ that

$$\Pr\left\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \le \hat{r}^2 + \gamma^2, \quad \boldsymbol{\beta} \ge \boldsymbol{0}\right\} = \alpha. \tag{7.9}$$

Therefore, if

$$\begin{aligned} \check{\phi}_k^{\text{lo}} &= \min_{\underline{\boldsymbol{\beta}}} \left\{\boldsymbol{w}_k^T \underline{\boldsymbol{\beta}} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) \le \hat{r}^2 + \gamma^2, \quad \underline{\boldsymbol{\beta}} \ge \boldsymbol{0}\right\} \\ \check{\phi}_k^{\text{up}} &= \max_{\underline{\boldsymbol{\beta}}} \left\{\boldsymbol{w}_k^T \underline{\boldsymbol{\beta}} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) \le \hat{r}^2 + \gamma^2, \quad \underline{\boldsymbol{\beta}} \ge \boldsymbol{0}\right\}, \end{aligned} \tag{7.10}$$

then for any number p of window vectors $\boldsymbol{w}_k$,

$$\Pr\left\{\check{\phi}_k^{\text{lo}} \le \boldsymbol{w}_k^T \boldsymbol{\beta} \le \check{\phi}_k^{\text{up}}, \quad k = 1, 2, \ldots, p\right\} = \alpha. \tag{7.11}$$

In Subsection 6.1 we saw that for full-rank problems with normally distributed errors, a $100\alpha\%$ confidence ellipsoid is obtained by choosing γ^2 to be the α point of the $\chi^2(n)$ distribution. With the addition of nonnegativity constraints, it becomes possible to estimate confidence intervals for underdetermined problems, but the choice of γ^2 is different. Rust and Burrus (1972, chap. 6) proved that if $\boldsymbol{Y}$ is normally distributed and

$$\hat{\boldsymbol{\beta}} = \left(\Sigma^{-1}\boldsymbol{X}\right)^{\dagger} \Sigma^{-1}\boldsymbol{Y} = \left(\boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X}\right)^{\dagger} \boldsymbol{X}^T \Sigma^{-2} \boldsymbol{Y} \tag{7.12}$$

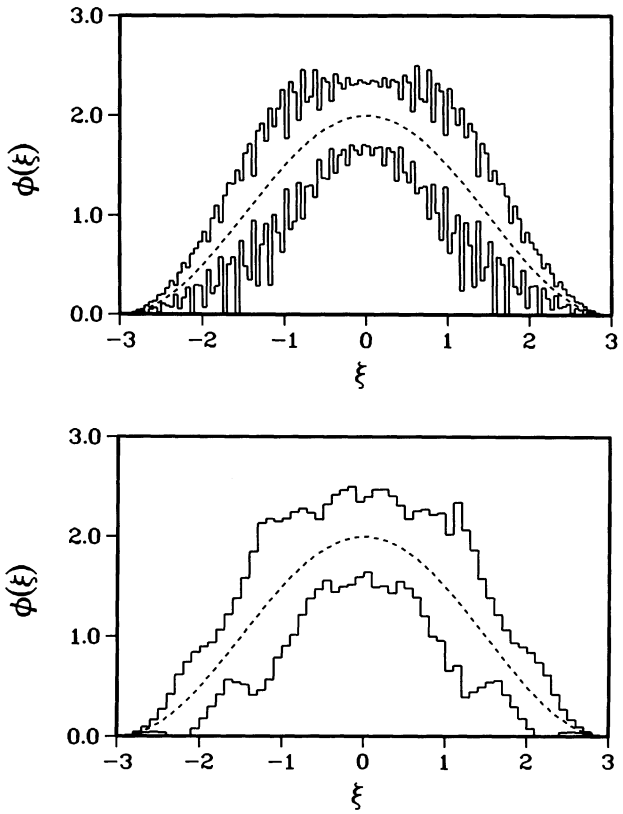

Figure 5. Nonnegatively Constrained, Simultaneous 95% Confidence Intervals From the Normal Distribution. The dashed lines are the exact solutions $\beta(\xi)$. Top: Problem I. The solid lines are histograms of lower and upper bounds for two-point averaging windows. Bottom: Problem II. The solid lines are histograms of lower and upper bounds for 3-point averaging windows.

is the corresponding generalized inverse solution vector, then $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \Sigma^{-2} \boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \sim \chi^2(\nu)$, where $\nu = \text{rank}(\boldsymbol{X})$. Therefore, if γ is taken to be the α point of the $\chi^2(\nu)$ distribution, then (7.10) defines the confidence intervals that can be computed with the BRACKET-LS algorithm (O'Leary and Rust 1986). Determining the exact rank ν is often a difficult and uncertain procedure, but it is better to overestimate the rank than to underestimate it. In the former case, the confidence intervals will be suboptimal (larger than necessary to assure the confidence level), but in the latter case they will be dishonest (too small to assure the claimed confidence level). A conservative procedure is to choose $\nu = \min\{m, n\}$.

Figure 5 gives plots of the simultaneous confidence intervals given by BRACKET-LS for Problem I using two-point averaging, and for Problem II using three-point averaging, with the measuring errors assumed to be normally distributed. The bounds for Problem I can be compared with those in the bottom frame of Figure 3 (p. 77) to assess the improvement obtained by using the nonnegativity constraints. The improvement is even more dramatic for Problem II because all intervals are unbounded for the unconstrained problem.

8. CONSTRAINED ONE-AT-A-TIME CONFIDENCE INTERVALS

8.1 THE BURRUS CONJECTURE

Nonnegatively constrained one-at-a-time confidence intervals cannot be defined in the same manner as constrained simultaneous intervals. More precisely, if κ is defined by (4.5), then valid $100\alpha\%$ one-at-a-time intervals cannot be calculated from (7.6) with $\mu^2 = \hat{r}^2 + \kappa^2$. In fact, it almost always happens that

$$\hat{r}^2 + \kappa^2 < \rho_{\min} = \min_{\underline{\beta} \geq \mathbf{0}} \left\{ (\boldsymbol{Y} - \boldsymbol{X}\underline{\beta})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\beta}) \right\}; \tag{8.1}$$

that is, the ellipsoid (5.4) had no points in common with the positive orthant. This difficulty arises because the regression models almost never have $m \gg n$. Accordingly, the classical least squares procedure produces a solution vector $\hat{\boldsymbol{\beta}}$ that models a large portion of the measurement errors $\boldsymbol{\epsilon}$ and hence gives an unrealistically low value for $\hat{r}^2$. This fact was pointed out by Burrus (1965) who also noted that the ellipsoid defined by

$$(\boldsymbol{Y} - \boldsymbol{X}\underline{\beta})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\beta}) \leq \rho_{\min} + \kappa^2 \tag{8.2}$$

always has a nonempty intersection with the positive orthant. He conjectured (Rust and Burrus 1972) that valid confidence intervals could be obtained by using that intersection as the constraint region; that is, by solving the problems

$$\begin{aligned} \check{\phi}^{\text{lo}} &= \min_{\underline{\beta}} \left\{ \boldsymbol{w}^T \underline{\beta} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\beta})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\beta}) = \rho_{\min} + \kappa^2, \quad \underline{\beta} \geq \mathbf{0} \right\} \\ \check{\phi}^{\text{up}} &= \max_{\underline{\beta}} \left\{ \boldsymbol{w}^T \underline{\beta} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\beta})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\beta}) = \rho_{\min} + \kappa^2, \quad \underline{\beta} \geq \mathbf{0} \right\}. \end{aligned} \tag{8.3}$$

We shall show in the following that this conjecture is indeed correct.

8.2 THE DUALITY THEOREM

To prove the Burrus Conjecture, it will be necessary to restate (8.3) in equivalent forms defined by the Duality Theorem for Nonlinear Programming (Wolfe 1961).

Wolfe's Duality Theorem: *Suppose $\boldsymbol{z}$ is an N vector of unknown variables, $f(\boldsymbol{z})$ is a scalar function, and $\boldsymbol{g}(\boldsymbol{z})$ is an M vector function of $\boldsymbol{z}$, and that $f(\boldsymbol{z})$ and $\boldsymbol{g}(\boldsymbol{z})$ are all convex and differentiable on an open set $\mathcal{Z}$. If*

$$\phi_P = \min_{\boldsymbol{z} \in \mathcal{Z}} \left\{ f(\boldsymbol{z}) \mid \boldsymbol{g}(\boldsymbol{z}) \geq \mathbf{0} \right\}, \tag{8.4}$$

and the constraints $\boldsymbol{g}(\boldsymbol{z}) \geq \mathbf{0}$ satisfy a constraint qualification, then $\phi_P = \phi_D$, where ϕ_D is determined by the dual problem

$$\phi_D = \max_{\{\boldsymbol{z}, \boldsymbol{v}\}} \left\{ f(\boldsymbol{z}) - \boldsymbol{g}^T \boldsymbol{v} \mid \nabla f(\boldsymbol{z}) = \left(\frac{\partial \boldsymbol{g}^T}{\partial \boldsymbol{z}} \right) \boldsymbol{v}, \quad \boldsymbol{v} \geq \mathbf{0} \right\}. \tag{8.5}$$

Equation (8.4) is called the primal problem. The constraint qualification can be any one of several regularity conditions on either the convexity or differentiability properties of the functions $g_i(\boldsymbol{z})$ (Mangasarian 1969, chaps. 5 and 7).

It is instructive to apply this theorem to the classical, full-rank interval estimation problems (5.1). To simplify the notation, let $\mu = +\sqrt{\hat{r}^2 + \kappa^2}$ and consider first the lower bound problem that can also be written with inequality constraints,

$$\hat{\phi}^{\text{lo}} = \min_{\underline{\boldsymbol{\beta}}} \left\{ \boldsymbol{w}^T \underline{\boldsymbol{\beta}} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) \leq \mu^2 \right\}, \tag{8.6}$$

because the indicated minimum is attained on the boundary of the ellipsoid. Many equivalent forms of the dual problem can be derived, but it is easiest to get the form we need by using the artifice of writing the problem in terms of the original variables $\underline{\boldsymbol{\beta}}$ and the scaled residual vector $\boldsymbol{r} = \Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}})$, which can also be specified by two vector inequalities,

$$\boldsymbol{r} - \Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) \geq \boldsymbol{0} \quad \text{and} \quad -\boldsymbol{r} + \Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) \geq \boldsymbol{0}. \tag{8.7}$$

Reformulating (8.6) in terms of the $(N = m + n)$ variables $\boldsymbol{z} \equiv \left(\underline{\boldsymbol{\beta}}^T, \boldsymbol{r}^T\right)^T$, gives

$$\boldsymbol{g}(\boldsymbol{z}) \equiv \begin{pmatrix} -\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) + \boldsymbol{r} \\ +\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) - \boldsymbol{r} \\ \mu - \sqrt{\boldsymbol{r}^T \boldsymbol{r}} \end{pmatrix} \geq \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ 0 \end{pmatrix} \tag{8.8}$$

for the constraints and

$$f(\boldsymbol{z}) \equiv \left(\boldsymbol{w}^T, \boldsymbol{0}^T\right) \boldsymbol{z} \tag{8.9}$$

for the objective function. The problem then becomes

$$\hat{\phi}^{\text{lo}} = \phi_P = \min_{\boldsymbol{z} \in \mathcal{Z}} \{ f(\boldsymbol{z}) \mid \boldsymbol{g}(\boldsymbol{z}) \geq \boldsymbol{0} \}, \tag{8.10}$$

where $\mathcal{Z} = \mathcal{R}^N$. The functions $f(\boldsymbol{z})$ and $\boldsymbol{g}(\boldsymbol{z})$ are convex and differentiable everywhere, and the constraints $\boldsymbol{g}(\boldsymbol{z}) \geq \boldsymbol{0}$ satisfy Slater's constraint qualification (Mangasarian 1969, sec. 5.4.3), so the Duality Theorem is applicable. Therefore

$$\hat{\phi}^{\text{lo}} = \phi_D = \max_{\{\boldsymbol{z}, \boldsymbol{v}\}} \left\{ f(\boldsymbol{z}) - \boldsymbol{g}^T \boldsymbol{v} \mid \nabla f(\boldsymbol{z}) = \left(\frac{\partial \boldsymbol{g}^T}{\partial \boldsymbol{z}}\right) \boldsymbol{v}, \quad \boldsymbol{v} \geq \boldsymbol{0} \right\}, \tag{8.11}$$

where $\boldsymbol{v}$ is a $(M = 2m + 1)$ vector of dual variables that can also be written

$$\boldsymbol{v}^T = \left(\boldsymbol{v}_1^T, \boldsymbol{v}_2^T, v_3\right)^T, \tag{8.12}$$

with subvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ of length m.

The derivatives required for the dual constraints are

$$\nabla f(\boldsymbol{z}) = \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{0} \end{pmatrix}, \left(\frac{\partial \boldsymbol{g}^T}{\partial \boldsymbol{z}}\right) = \begin{pmatrix} \boldsymbol{X}^T \Sigma^{-1} & -\boldsymbol{X}^T \Sigma^{-1} & \boldsymbol{0} \\ \boldsymbol{I}_m & -\boldsymbol{I}_m & -\frac{\boldsymbol{r}}{\sqrt{\boldsymbol{r}^T \boldsymbol{r}}} \end{pmatrix}, \tag{8.13}$$

so the constraints themselves can be written

$$\boldsymbol{X}^T \Sigma^{-1}(\boldsymbol{v}_1 - \boldsymbol{v}_2) = \boldsymbol{w}, \quad (\boldsymbol{v}_1 - \boldsymbol{v}_2) - v_3 \frac{\boldsymbol{r}}{\sqrt{\boldsymbol{r}^T \boldsymbol{r}}} = \boldsymbol{0}, \quad \boldsymbol{v}_1 \geq \boldsymbol{0}, \ \boldsymbol{v}_2 \geq \boldsymbol{0}, \ v_3 \geq 0. \tag{8.14}$$

Note that the dual vectors v_1 and v_2 appear only as components of the difference $v_1 - v_2$. Even though v_1 and v_2 are constrained to be nonnegative, their difference is totally unconstrained. Defining new, unconstrained dual variables $v_0 \equiv v_1 - v_2$ allows the constraints to be written

$$X^T\Sigma^{-1}v_0 = w, \quad v_0 = v_3\frac{r}{\sqrt{r^Tr}}, \quad v_3 \geq 0 \tag{8.15}$$

and then eliminating the variables v_0 gives

$$v_3\frac{X^T\Sigma^{-1}r}{\sqrt{r^Tr}} = w, \quad v_3 \geq 0. \tag{8.16}$$

Using the definition of v_0 and Equations (8.15) and (8.16), the objective function for the dual problem can be written

$$\begin{aligned} f(z) - g^T(z)v &= w^T\underline{\beta} + \left[Y^T\Sigma^{-1} - \underline{\beta}^TX^T\Sigma^{-1} - r^T\right](v_1 - v_2) \\ &\quad + v_3\left(\sqrt{r^Tr} - \mu\right) \end{aligned} \tag{8.17}$$

$$\begin{aligned} &= w^T\underline{\beta} + \left[Y^T\Sigma^{-1} - \underline{\beta}^TX^T\Sigma^{-1} - r^T\right]v_0 \\ &\quad + v_3\left(\sqrt{r^Tr} - \mu\right) \end{aligned} \tag{8.18}$$

$$\begin{aligned} &= v_3\frac{r^T\Sigma^{-1}X\underline{\beta}}{\sqrt{r^Tr}} + \left[Y^T\Sigma^{-1} - \underline{\beta}^TX^T\Sigma^{-1} - r^T\right]\frac{v_3r}{\sqrt{r^Tr}} \\ &\quad + v_3\left(\sqrt{r^Tr} - \mu\right) \end{aligned} \tag{8.19}$$

$$= v_3\left(\frac{Y^T\Sigma^{-1}r}{\sqrt{r^Tr}} - \mu\right). \tag{8.20}$$

To simplify, we introduce a change of variables

$$u \equiv \frac{v_3\Sigma^{-1}r}{\sqrt{r^Tr}}, \tag{8.21}$$

and note that by (8.16) u is constrained only by $X^Tu = w$. Multiplying (8.21) by Σ and "squaring" yields $u^T\Sigma^2u = v_3^2$, so the dual objective function becomes

$$f(z) - g^T(z)v = Y^Tu - \mu\sqrt{u^T\Sigma^2u}, \tag{8.22}$$

which yields, after substituting $\mu = \sqrt{\hat{r}^2 + \kappa^2}$, the final form of the dual problem—that is,

$$\hat{\phi}^{\text{lo}} = \max_u\left\{u^TY - \left(\hat{r}^2 + \kappa^2\right)^{\frac{1}{2}}\sqrt{u^T\Sigma^2u} \mid u^TX = w^T\right\}. \tag{8.23}$$

In a similar manner, it can be shown that the dual problem for the upper bound problem in (5.1) can be written

$$\hat{\phi}^{\text{up}} = \min_u\left\{u^TY + \left(\hat{r}^2 + \kappa^2\right)^{\frac{1}{2}}\sqrt{u^T\Sigma^2u} \mid u^TX = w^T\right\}. \tag{8.24}$$

It is easy to independently verify that these two dual problems are equivalent to the primals (5.1) because they too can be solved by Lagrange multipliers. The solution vectors are

$$\hat{\boldsymbol{u}}^{\text{lo}} = \hat{\boldsymbol{u}} + \frac{\sqrt{\hat{\boldsymbol{u}}^T\Sigma^2\hat{\boldsymbol{u}}}}{\kappa}\Sigma^{-2}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

$$\hat{\boldsymbol{u}}^{\text{up}} = \hat{\boldsymbol{u}} - \frac{\sqrt{\hat{\boldsymbol{u}}^T\Sigma^2\hat{\boldsymbol{u}}}}{\kappa}\Sigma^{-2}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{8.25}$$

where $\hat{\boldsymbol{u}}$ is given by (3.8) and $\hat{\boldsymbol{\beta}}$ by (3.9). The corresponding optimal values are

$$\begin{aligned}\hat{\phi}^{\text{lo}} &= \left(\hat{\boldsymbol{u}}^{\text{lo}}\right)^T\boldsymbol{Y} - \left(\hat{r}^2 + \kappa^2\right)^{\frac{1}{2}}\sqrt{\left(\hat{\boldsymbol{u}}^{\text{lo}}\right)^T\Sigma^2\hat{\boldsymbol{u}}^{\text{lo}}} = \hat{\boldsymbol{u}}^T\boldsymbol{Y} - \kappa\sqrt{\hat{\boldsymbol{u}}^T\Sigma^2\hat{\boldsymbol{u}}},\\ \hat{\phi}^{\text{up}} &= \left(\hat{\boldsymbol{u}}^{\text{up}}\right)^T\boldsymbol{Y} + \left(\hat{r}^2 + \kappa^2\right)^{\frac{1}{2}}\sqrt{\left(\hat{\boldsymbol{u}}^{\text{up}}\right)^T\Sigma^2\hat{\boldsymbol{u}}^{\text{up}}} = \hat{\boldsymbol{u}}^T\boldsymbol{Y} + \kappa\sqrt{\hat{\boldsymbol{u}}^T\Sigma^2\hat{\boldsymbol{u}}},\end{aligned} \tag{8.26}$$

which are the same as the values given by (4.4).

The primal problems (5.1) correspond to a Gaussian formulation of linear interval estimation and the dual problems (8.23) and (8.24) correspond to the Markov formulation. In the objective functions of the latter, the inner product $\boldsymbol{u}^T\boldsymbol{Y}$ is a linear estimator of the unknown ϕ^* and the quantity $\boldsymbol{u}^T\Sigma^{-2}\boldsymbol{u}$ is the variance of that estimator. The constraint functions common to both problems require the estimators to be unbiased. When nonnegativity constraints are imposed on the problems, the forms of the objective functions are unaffected, but the constraint functions change in an interesting way.

8.3 Lower-Biased and Upper-Biased Estimators

Now consider the nonnegatively constrained interval estimation problems (8.3). The lower bound primal problem is

$$\check{\phi}^{\text{lo}} = \min_{\underline{\boldsymbol{\beta}}}\left\{\boldsymbol{w}^T\underline{\boldsymbol{\beta}} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}})^T\Sigma^{-2}(\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) \le \mu^2, \quad \underline{\boldsymbol{\beta}} \ge \boldsymbol{0}\right\}, \tag{8.27}$$

where $\mu = \sqrt{\rho_{\min} + \kappa^2}$. The transformation of this problem to its dual form is similar to that given in the preceding section for the unconstrained problem (8.6). The only difference is the additional set of n constraints, $\underline{\boldsymbol{\beta}} \ge \boldsymbol{0}$, that augment the primal constraint relations (8.8). The dual vector (8.12) is also augmented by n additional variables $\boldsymbol{v}_4$ to give $\boldsymbol{v}^T = \left(\boldsymbol{v}_1^T, \boldsymbol{v}_2^T, v_3, \boldsymbol{v}_4^T\right)^T$, and the constraint expressions (8.14) are changed to

$$\boldsymbol{X}^T\Sigma^{-1}(\boldsymbol{v}_1 - \boldsymbol{v}_2) + \boldsymbol{v}_4 = \boldsymbol{w}, \ (\boldsymbol{v}_1 - \boldsymbol{v}_2) - v_3\frac{\boldsymbol{r}}{\sqrt{\boldsymbol{r}^T\boldsymbol{r}}} = \boldsymbol{0},$$

$$\boldsymbol{v}_1 \ge \boldsymbol{0}, \ \boldsymbol{v}_2 \ge \boldsymbol{0}, \ v_3 \ge 0, \ \boldsymbol{v}_4 \ge \boldsymbol{0}. \tag{8.28}$$

The objective function remains the same as it was for the classical problem, and the definition of $\boldsymbol{v}_0$ and the determination of v_3 are the same also. The net result is that

$$\check{\phi}^{\text{lo}} = \max_{\{\boldsymbol{v}_0, \boldsymbol{v}_4\}}\left\{\boldsymbol{Y}^T\Sigma^{-1}\boldsymbol{v}_0 - \mu\sqrt{\boldsymbol{v}_0^T\boldsymbol{v}_0} \mid \boldsymbol{X}^T\Sigma^{-1}\boldsymbol{v}_0 + \boldsymbol{v}_4 = \boldsymbol{w}, \quad \boldsymbol{v}_4 \ge \boldsymbol{0}\right\}. \tag{8.29}$$

The objective function does not depend on v_4, so the constraints can be written $X^T\Sigma^{-1}v_0 \leq w$. Defining u as in (8.21) and restoring the value of μ gives

$$\check{\phi}^{\text{lo}} = \max_{u}\left\{u^TY - \left(\rho_{\min} + \kappa^2\right)^{\frac{1}{2}}\sqrt{u^T\Sigma^2 u} \mid u^TX \leq w^T\right\}. \tag{8.30}$$

Similarly, the dual statement of the upper bound problem can be shown to be

$$\check{\phi}^{\text{up}} = \min_{u}\left\{u^TY + \left(\rho_{\min} + \kappa^2\right)^{\frac{1}{2}}\sqrt{u^T\Sigma^2 u} \mid u^TX \geq w^T\right\}. \tag{8.31}$$

The objective functions for the current dual problems differ from those of the classical duals (8.23) and (8.24) only in the substitution of $\rho_{\min}$ for $\hat{r}^2$. The constraints, however, are quite different. In the classical case, the estimators u^TY are required to be unbiased in both cases. For the nonnegatively constrained problems, the estimator for the lower bound is required to be lower biased, and the estimator for the upper bound is required to be upper biased. This equivalence of bias constraints in the Markov formulation to nonnegativity constraints in the Gauss formulation is a surprising result whose significance we are at present unable to assess. It seems very logical to use lower-biased estimators when seeking a lower bound and upper-biased estimators when seeking an upper bound, but such constraints appear at first glance to be less restrictive than the classical unbiasedness constraints. In practical problems, nonnegatively constrained confidence intervals are smaller than the corresponding classical intervals, even though the substitution of $\rho_{\min}$ for $\hat{r}^2$ in the objective functions should tend to increase the size of the former. It would seem then that the upper and lower bias constraints are considerably stronger than the unbiasedness constraint.

8.4 PROOF OF THE BURRUS CONJECTURE

Having established the equivalence of the dual problems (8.30) and (8.31) to the primals (8.3), it is not difficult to prove the Burrus Conjecture, which we now state as a theorem.

Theorem. *Let Y be a given m vector, X a given $m \times n$ matrix, and β an unknown n vector satisfying*

$$Y = X\beta + \epsilon, \quad \epsilon \sim N\left(0, \Sigma^2\right), \quad \Pr\{\beta \geq 0\} = 1. \tag{8.32}$$

Let $\rho_{\min}$ be defined by

$$\rho_{\min} = \min_{\underline{\beta}}\left\{(Y - X\underline{\beta})^T\Sigma^{-2}(Y - X\underline{\beta}) \mid \underline{\beta} \geq 0\right\}. \tag{8.33}$$

Let α be a given probability $(0 < \alpha < 1)$, and suppose that the value κ is chosen so that

$$\frac{1}{\sqrt{2\pi}}\int_{-\kappa}^{+\kappa}\exp\left(-\frac{\eta^2}{2}\right)d\eta = \alpha. \tag{8.34}$$

If w is a given n vector, and

$$\check{\phi}^{\text{lo}} = \min_{\underline{\beta}}\left\{w^T\underline{\beta} \mid (Y - X\underline{\beta})^T\Sigma^{-2}(Y - X\underline{\beta}) \leq \rho_{\min} + \kappa^2, \quad \underline{\beta} \geq 0\right\}$$

$$\check{\phi}^{\text{up}} = \max_{\underline{\boldsymbol{\beta}}} \left\{ \boldsymbol{w}^T \underline{\boldsymbol{\beta}} \mid (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}})^T \Sigma^{-2} (\boldsymbol{Y} - \boldsymbol{X}\underline{\boldsymbol{\beta}}) \leq \rho_{\min} + \kappa^2, \quad \underline{\boldsymbol{\beta}} \geq \boldsymbol{0} \right\}, \tag{8.35}$$

then

$$\Pr\left\{ \check{\phi}^{\text{lo}} \leq \boldsymbol{w}^T \boldsymbol{\beta} \leq \check{\phi}^{\text{up}} \right\} \geq \alpha. \tag{8.36}$$

Proof: For any m vector $\boldsymbol{u}$, the reduced random variable

$$\eta = \frac{\boldsymbol{u}^T \boldsymbol{Y} - \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{\beta}}{\sqrt{\boldsymbol{u}^T \Sigma^2 \boldsymbol{u}}} \tag{8.37}$$

is distributed as $n(0,1)$. It follows that

$$\Pr\left\{ \left(\boldsymbol{u}^T \boldsymbol{Y} - \kappa \sqrt{\boldsymbol{u}^T \Sigma^2 \boldsymbol{u}} \right) \leq \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{\beta} \leq \left(\boldsymbol{u}^T \boldsymbol{Y} + \kappa \sqrt{\boldsymbol{u}^T \Sigma^2 \boldsymbol{u}} \right) \right\} = \alpha. \tag{8.38}$$

Define constants θ_1 and θ_2 by

$$\begin{aligned} &\Pr\left\{ \left(\boldsymbol{u}^T \boldsymbol{Y} - \kappa \sqrt{\boldsymbol{u}^T \Sigma^2 \boldsymbol{u}} \right) \leq \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{\beta} \right\} \\ &\qquad = \frac{1}{\sqrt{2\pi}} \int_{-\kappa}^{\infty} \exp\left(-\frac{\eta^2}{2} \right) d\eta = \alpha + \theta_2 \\ &\Pr\left\{ \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{\beta} \leq \left(\boldsymbol{u}^T \boldsymbol{Y} + \kappa \sqrt{\boldsymbol{u}^T \Sigma^2 \boldsymbol{u}} \right) \right\} \\ &\qquad = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\kappa} \exp\left(-\frac{\eta^2}{2} \right) d\eta = \theta_1 + \alpha, \end{aligned} \tag{8.39}$$

and note that

$$\theta_1 + \alpha + \theta_2 = 1. \tag{8.40}$$

We saw in the preceding section that $\check{\phi}^{\text{lo}}$ and $\check{\phi}^{\text{up}}$ can also be defined as

$$\begin{aligned} \check{\phi}^{\text{lo}} &= \max_{\boldsymbol{u}} \left\{ \boldsymbol{u}^T \boldsymbol{Y} - \left(\rho_{\min} + \kappa^2 \right)^{\frac{1}{2}} \sqrt{\boldsymbol{u}^T \Sigma^2 \boldsymbol{u}} \mid \boldsymbol{u}^T \boldsymbol{X} \leq \boldsymbol{w}^T \right\} \\ \check{\phi}^{\text{up}} &= \min_{\boldsymbol{u}} \left\{ \boldsymbol{u}^T \boldsymbol{Y} + \left(\rho_{\min} + \kappa^2 \right)^{\frac{1}{2}} \sqrt{\boldsymbol{u}^T \Sigma^2 \boldsymbol{u}} \mid \boldsymbol{u}^T \boldsymbol{X} \geq \boldsymbol{w}^T \right\}. \end{aligned} \tag{8.41}$$

Consider first the upper bound problem and let $\boldsymbol{u}_{\text{up}}$ be its solution vector—that is,

$$\check{\phi}^{\text{up}} = \boldsymbol{u}_{\text{up}}^T \boldsymbol{Y} + \left(\rho_{\min} + \kappa^2 \right)^{\frac{1}{2}} \sqrt{\boldsymbol{u}_{\text{up}}^T \Sigma^2 \boldsymbol{u}_{\text{up}}}, \tag{8.42}$$

with

$$\boldsymbol{u}_{\text{up}}^T \boldsymbol{X} \geq \boldsymbol{w}^T. \tag{8.43}$$

From the second relation in (8.39), because $\kappa \leq \sqrt{\rho_{\min} + \kappa^2}$, it follows that

$$\Pr\left\{ \boldsymbol{u}_{\text{up}}^T \boldsymbol{X} \boldsymbol{\beta} \leq \left[\boldsymbol{u}_{\text{up}}^T \boldsymbol{Y} + \left(\rho_{\min} + \kappa^2 \right)^{\frac{1}{2}} \sqrt{\boldsymbol{u}_{\text{up}}^T \Sigma^2 \boldsymbol{u}_{\text{up}}} \right] \right\} \geq \theta_1 + \alpha, \tag{8.44}$$

or, by (8.42), that

$$\Pr\left\{\boldsymbol{u}_{\text{up}}^T \boldsymbol{X}\boldsymbol{\beta} \leq \check{\phi}^{\text{up}}\right\} \geq \theta_1 + \alpha. \tag{8.45}$$

Because $\boldsymbol{\beta} \geq \mathbf{0}$, it follows from (8.43) that $\boldsymbol{w}^T\boldsymbol{\beta} \leq \boldsymbol{u}_{\text{up}}^T \boldsymbol{X}\boldsymbol{\beta}$. Combining this inequality with (8.45) gives

$$\Pr\left\{\boldsymbol{w}^T\boldsymbol{\beta} \leq \check{\phi}^{\text{up}}\right\} \geq \theta_1 + \alpha. \tag{8.46}$$

In a similar manner, it is easy to show that

$$\Pr\left\{\check{\phi}^{\text{lo}} \leq \boldsymbol{w}^T\boldsymbol{\beta}\right\} \geq \alpha + \theta_2. \tag{8.47}$$

Now,

$$\begin{aligned}\Pr\left\{\check{\phi}^{\text{lo}} \leq \boldsymbol{w}^T\boldsymbol{\beta} \leq \check{\phi}^{\text{up}}\right\} &= \Pr\left[\left\{\check{\phi}^{\text{lo}} \leq \boldsymbol{w}^T\boldsymbol{\beta}\right\} \bigcap \left\{\boldsymbol{w}^T\boldsymbol{\beta} \leq \check{\phi}^{\text{up}}\right\}\right] \\ &= \Pr\left\{\check{\phi}^{\text{lo}} \leq \boldsymbol{w}^T\boldsymbol{\beta}\right\} + \Pr\left\{\boldsymbol{w}^T\boldsymbol{\beta} \leq \check{\phi}^{\text{up}}\right\} \\ &\quad - \Pr\left[\left\{\check{\phi}^{\text{lo}} \leq \boldsymbol{w}^T\boldsymbol{\beta}\right\} \bigcup \left\{\boldsymbol{w}^T\boldsymbol{\beta} \leq \check{\phi}^{\text{up}}\right\}\right], \end{aligned} \tag{8.48}$$

so by (8.46) and (8.47),

$$\begin{aligned}\Pr\left\{\check{\phi}^{\text{lo}} \leq \boldsymbol{w}^T\boldsymbol{\beta} \leq \check{\phi}^{\text{up}}\right\} &\geq \theta_1 + 2\alpha + \theta_2 \\ &\quad - \Pr\left[\left\{\check{\phi}^{\text{lo}} \leq \boldsymbol{w}^T\boldsymbol{\beta}\right\} \bigcup \left\{\boldsymbol{w}^T\boldsymbol{\beta} \leq \check{\phi}^{\text{up}}\right\}\right]. \end{aligned} \tag{8.49}$$

From definition (8.35), we have $\check{\phi}^{\text{lo}} < \check{\phi}^{\text{up}}$, so

$$\Pr\left[\left\{\check{\phi}^{\text{lo}} \leq \boldsymbol{w}^T\boldsymbol{\beta}\right\} \bigcup \left\{\boldsymbol{w}^T\boldsymbol{\beta} \leq \check{\phi}^{\text{up}}\right\}\right] = 1. \tag{8.50}$$

Substituting this result into (8.49) and using (8.40) gives

$$\Pr\left\{\check{\phi}^{\text{lo}} \leq \boldsymbol{w}^T\boldsymbol{\beta} \leq \check{\phi}^{\text{up}}\right\} \geq \theta_1 + 2\alpha + \theta_2 - 1 = \alpha, \tag{8.51}$$

which completes the proof of the theorem. □

Note that the theorem makes no restrictions on m, n, or $\text{rank}(\boldsymbol{X})$. For an underdetermined system in which $\boldsymbol{X}$ has some negative elements, it is possible, but not very probable, that one of the two problems (8.35) may be unbounded. In that case, the theorem is still true, but the confidence interval is semi-infinite.

In proving the theorem we assumed that the errors $\boldsymbol{\epsilon}$ were normally distributed. This is generally not a serious limitation in practice, but it is worth noting that the theorem can be proved for a wider class of possible error distributions. The essential restriction required by the proof is that the reduced random variable η have the same probability distribution for all m vectors $\boldsymbol{u}$.

Corollary 1. *Let $\boldsymbol{Y}$ be a given m vector, $\boldsymbol{X}$ a given $m \times n$ matrix, and $\boldsymbol{\beta}$ an unknown n vector satisfying*

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathcal{E}(\boldsymbol{\epsilon}) = \mathbf{0}, \quad \mathcal{E}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right) = \Sigma^2, \quad \Pr\{\boldsymbol{\beta} \geq \mathbf{0}\} = 1. \tag{8.52}$$

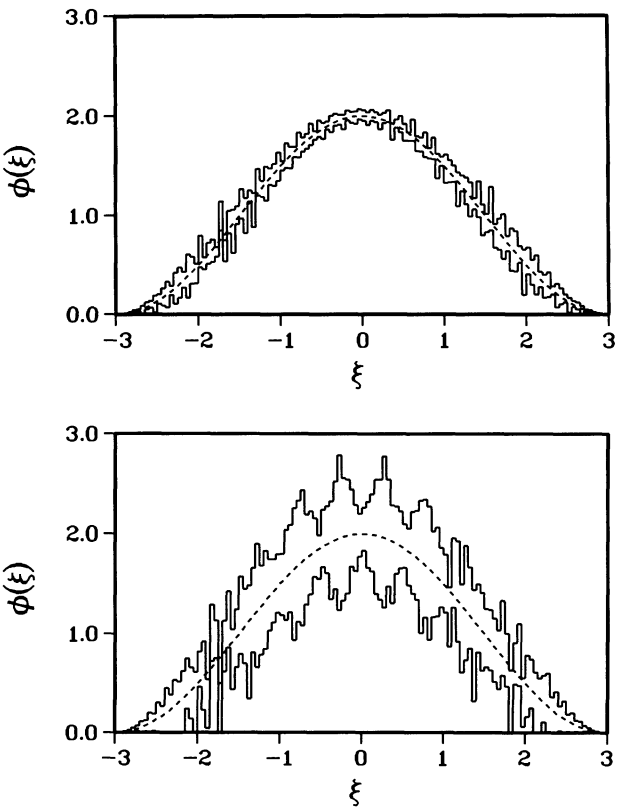

Figure 6. Nonnegatively Constrained, One-at-a-Time 95% Confidence Intervals From the Normal Distribution. The dashed lines are the exact solutions $\beta(\xi)$. Top: Problem I. The solid lines are histograms of lower and upper bounds for two-point averaging windows. Bottom: Problem II. The solid lines are histograms of lower and upper bounds for two-point averaging windows.

Let $\rho_{\min}$ be defined by (8.33) and assume that the errors ϵ are distributed in such a manner that the reduced random variable

$$\eta = \frac{\boldsymbol{u}^T\boldsymbol{Y} - \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{\beta}}{\sqrt{\boldsymbol{u}^T\Sigma^2\boldsymbol{u}}} \tag{8.53}$$

has the same probability density function $f(\eta)$ for all m vectors $\boldsymbol{u}$. Let α be a given probability $(0 < \alpha < 1)$, and suppose that the value κ is chosen so that

$$\Pr\{-\kappa \le \eta \le \kappa\} = \int_{-\kappa}^{\kappa} f(\eta)\, d\eta = \alpha. \tag{8.54}$$

If $\boldsymbol{w}$ is a given n vector, and if $\check{\phi}^{\text{lo}}$ and $\check{\phi}^{\text{up}}$ are defined by (8.35), then

$$\Pr\left\{\check{\phi}^{\text{lo}} \le \boldsymbol{w}^T\boldsymbol{\beta} \le \check{\phi}^{\text{up}}\right\} \ge \alpha. \tag{8.55}$$

Proof: The proof differs from the proof of the theorem only in the definitions of

θ_1 and θ_2—that is, (8.39) is replaced by

$$\begin{aligned} \Pr\left\{\left(\boldsymbol{u}^T\boldsymbol{Y} - \kappa\sqrt{\boldsymbol{u}^T\Sigma^2\boldsymbol{u}}\right) \le \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{\beta}\right\} &= \int_{-\kappa}^{\infty} f(\eta)d\eta = \alpha + \theta_2 \\ \Pr\left\{\boldsymbol{u}^T\boldsymbol{X}\boldsymbol{\beta} \le \left(\boldsymbol{u}^T\boldsymbol{Y} + \kappa\sqrt{\boldsymbol{u}^T\Sigma^2\boldsymbol{u}}\right)\right\} &= \int_{-\infty}^{+\kappa} f(\eta)\,d\eta = \theta_1 + \alpha. \end{aligned} \quad (8.56)$$

The values θ_1 and θ_2 satisfy $\theta_1 + \alpha + \theta_2 = 1$, and the remainder of the proof follows exactly as before. □

8.5 EXAMPLES

Figure 6 gives plots of the nonnegatively constrained 95% confidence intervals for both Problems I and II, using two-point averaging. The Problem I bounds can be compared with those in the top frame of Figure 3 (p. 77) to assess the improvement to be attributed to the nonnegativity constraints. Problem I is not a bad problem, so the improvement is less striking than that normally attained for real-world problems. The bounds for Problem II do represent a striking improvement because, in the absence of the nonnegativity constraints, all the confidence intervals are unbounded. The bounds for Problem II could be significantly improved by using three-point averaging.

[*Received January 1992. Revised July 1993.*]

REFERENCES

Burrus, W. R. (1965), "Utilization of A Priori Information by Means of Mathematical Programming in the Statistical Interpretation of Measured Distributions," Technical Report ORNL-3743, Oak Ridge National Laboratory, Ph.D. Thesis, Ohio State University, Physics Dept.

Burrus, W. R., Rust, B. W., and Cope, J. E. (1980), "Constrained Interval Estimation for Linear Models with Ill-Conditioned Equations," in *Information Linkage between Applied Mathematics and Industry II*, ed. A. L. Schoenstadt, F. D. Faulkner, R. Frank, and I. B. Russak, New York: Academic Press, pp. 1–38.

Cope, J. E., and Rust, B. W. (1979), "Bounds on Solutions of Linear Systems with Inaccurate Data," *SIAM Journal on Numerical Analysis*, 16, 950–963.

Graybill, F. A. (1976), *Theory and Application of the Linear Model*, North Scituate, MA: Duxbury Press.

Hogg, R. V., and Craig, A. T. (1965), *Introduction to Mathematical Statistics*, New York: Macmillan.

Hunter, D. (1976), "An Upper Bound for the Probability of a Union," *Journal of Applied Probability*, 13, 597–603.

Lawson, C. L., and Hanson, R. J. (1974), *Solving Least Squares Problems*, Englewood Cliffs, NJ: Prentice Hall.

Mangasarian, O. L. (1969), *Nonlinear Programming*, New York: McGraw-Hill.

O'Leary, D. P., and Rust, B. W. (1986), "Confidence Intervals for Inequality-Constrained Least Squares Problems, With Applications to Ill-Posed Problems," *SIAM Journal on Scientific and Statistical Computing*, 7, 473–489.

Olkin, I., and Pratt, J. W. (1958), "A Multivariate Tchebycheff Inequality," *Annals of Mathematical Statistics*, 29, 226–234.

Phillips, D. L. (1962), "A Technique for the Numerical Solution of Certain Integral Equations of the First Kind," *Journal Assoc. Comput. Mach.*, 9, 84–97.

Pierce, J. E., and Rust, B. W. (1985), "Constrained Least Squares Interval Estimation," *SIAM Journal on Scientific and Statistical Computing*, 6, 670–683.

Rust, B. W., and Burrus, W. R. (1971), "Suboptimal Methods for Solving Constrained Estimation Problems," Technical Report DASA 2604, Tennecomp Systems, Inc., Oak Ridge, Tennessee.

——— (1972), *Mathematical Programming and the Numerical Solution of Linear Equations*, New York: American Elsevier.

Stoline, M. R. (1983), "The Hunter Method of Simultaneous Inference and Its Recommended Use for Applications Having Large Known Correlation Structures," *Journal of the American Statistical Association*, 78, 366–370.

Tarantola, A. (1987), *Inverse Problem Theory*, Amsterdam: Elsevier.

Tong, Y. L. (1980), *Probability Inequalities in Multivariate Distributions*, New York: Academic Press.

Wolfe, P. (1961), "A Duality Theorem for Non-Linear Programming," *Quarterly of Applied Mathematics*, 19, 239–244.

Worsley, K. J. (1982), "An Improved Bonferroni Inequality and Applications," *Biometrika*, 69, 297–302.