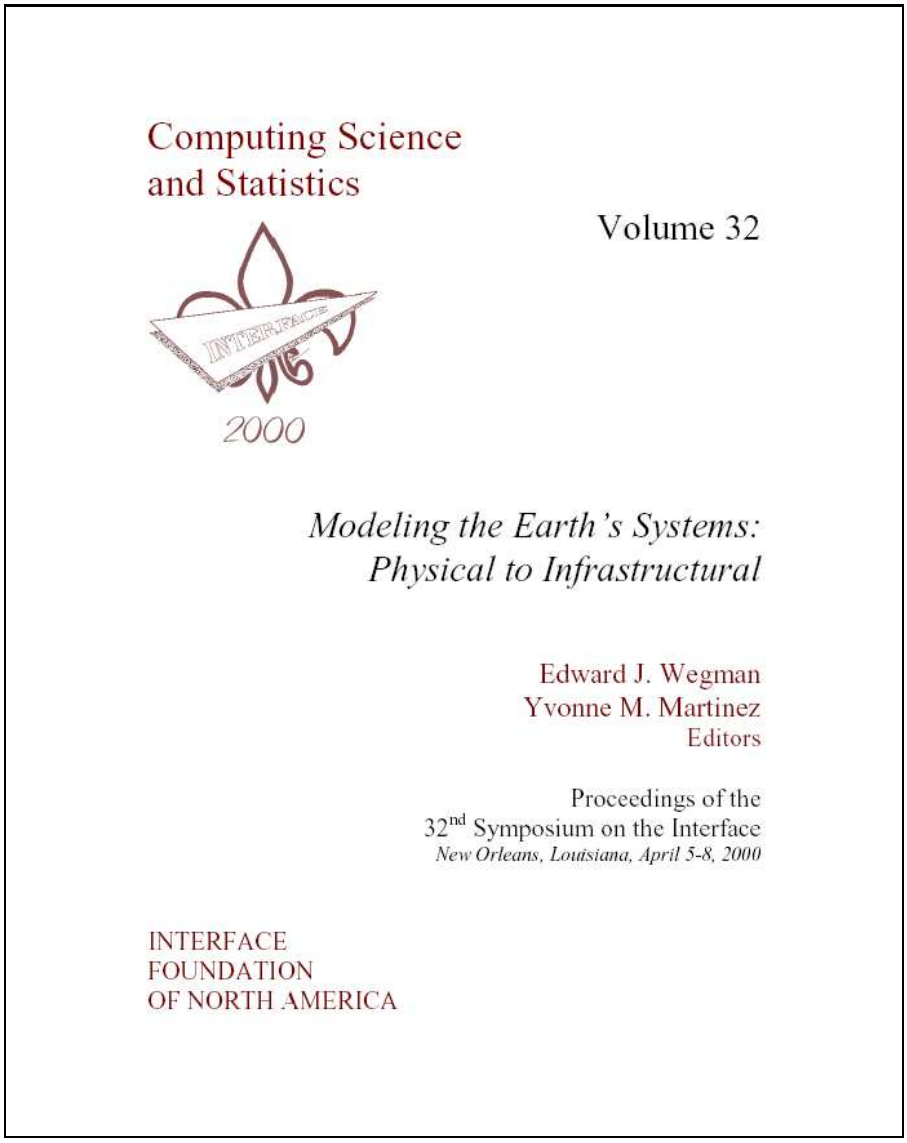
Parameter Selection for Constrained Solutions to Ill-Posed Problems

Bert W. Rust

Reprinted from the CD



Rust, B. W. (2000) "Parameter Selection for Constrained Solutions to Ill-Posed Problems," *Computing Science and Statistics*, **32**, 333-347.

– or –

Rust, B. W. (2000) "Parameter Selection for Constrained Solutions to Ill-Posed Problems," *Computing Science and Statistics*, **32**, /I2000Proceedings/BRust/BRust.pdf.

Parameter Selection for Constrained Solutions to Ill-Posed Problems

Bert W. Rust

National Institute of Standards and Technology
100 Bureau Drive Stop 8910
Gaithersburg, MD 20899-8910
bwr@cam.nist.gov

Abstract

Many physical measurements $y(t_i)$ can be modelled by a system of linear, first kind integral equations

$$y(t_i) = \int_{\xi_{lo}}^{\xi_{up}} K(t_i,\xi)x(\xi)d\xi + \epsilon_i \ , \quad i = 1, 2, \ldots, m \ , \quad (1)$$

where $x(\xi)$ is the function being measured, the $K(t_i,\xi)$ are known instrument response functions, and the ϵ_i are random measuring errors. Discretizing the integrals produces an ill-conditioned linear regression model, and minimizing the sum of squared residuals forces variance that should properly be relegated to the residuals into the least squares estimate which becomes a wildly oscillating, spurious approximation to $x(\xi)$. Adopting the principle that acceptable residuals should resemble the ϵ_i leads to three statistical tests for the suitability of an estimate. Stabilized estimates can be obtained either by appending a set of regularization constraints to the model or by truncating the singular value decomposition of the matrix. Using a test problem with known solution, it is shown that conventional methods for choosing the regularization parameter yield unacceptable estimates, but that the three statistical tests can be used to choose an optimal value. It is also shown that truncating the distribution of singular values does not work as well as discarding the components of the rotated data vector that are overwhelmed by measurement errors, and that the three statistical tests can be used to optimize the choice of the truncation threshold.

1 Introduction

When a measuring instrument is used to observe a function $x(\xi)$ on some interval $\xi_{lo} \le \xi \le \xi_{up}$, the resulting measurements can often be modelled by a system of linear, first kind integral equations (1) where the $y_i \equiv y(t_i)$ are the measured values, the $K_i(\xi) \equiv K(t_i,\xi)$ are known instrument response functions, and the ϵ_i are random measuring errors. The exactly known t_i are discrete values of a physical variable t which is an alias for ξ, so the y_i comprise a discrete approximation to a function $y(t)$ which is a smoothed (and possibly distorted) approximation to $x(\xi)$.

Discretizing the integrals in (1) leads to a usually poorly conditioned linear regression model

$$\mathbf{y} = \mathbf{K}\mathbf{x}^* + \boldsymbol{\epsilon} \ , \quad \mathcal{E}(\boldsymbol{\epsilon}) = \mathbf{0} \ , \quad \mathcal{E}\left(\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}^T\right) = \mathbf{S}^2 \ , \qquad (2)$$

where $\mathbf{x}^*$ is a discrete n-vector approximation to the function $x(\xi)$, with $n \le m$, and $\mathbf{S}^2$ is a positive definite variance matrix for the measurement errors. It will be assumed here that the errors are independently normally distributed with known variances, i.e., that

$$\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \mathbf{S}^2\right) \quad . \quad \mathbf{S}^2 = \mathrm{diag}\left(s_1^2, s_2^2, \cdots, s_m^2\right) \quad , \qquad (3)$$

where good estimates are available for the s_i. This allows the problem to be scaled by the tranformations

$$\mathbf{b} \equiv \mathbf{S}^{-1}\mathbf{y} \ , \quad \mathbf{A} \equiv \mathbf{S}^{-1}\mathbf{K} \ , \quad \boldsymbol{\eta} \equiv \mathbf{S}^{-1}\boldsymbol{\epsilon} \qquad (4)$$

to give

$$\mathbf{b} = \mathbf{A}\mathbf{x}^* + \boldsymbol{\eta} \ , \quad \boldsymbol{\eta} \sim N(\ \mathbf{0}\ ,\ \mathbf{I}_m\ ) \ , \qquad (5)$$

where $\mathbf{I}_m$ is m-order identity matrix. Note that in cases where the measurement errors are correlated, i.e., when $\mathbf{S}^2$ is not a diagonal matrix, the same scaling can be obtained by substituting for the matrix $\mathbf{S}$ in (4) the lower triangular matrix $\mathbf{L}$ obtained from the Cholesky factorization $\mathbf{S}^2 = \mathbf{L}\mathbf{L}^T$.

2 Estimates

Let $\hat{\mathbf{x}}$ be any estimate of the solution vector $\mathbf{x}^*$ and

$$\hat{\mathbf{r}} \ \equiv \ \mathbf{b} - \mathbf{A}\hat{\mathbf{x}} \qquad (6)$$

be the corresponding residual vector. It is instructive to rewrite the scaled model (5) in the form

$$\boldsymbol{\eta} \ = \ \mathbf{b} - \mathbf{A}\mathbf{x}^* \ , \quad \boldsymbol{\eta} \sim N(\ \mathbf{0}\ ,\ \mathbf{I}_m\ ) \ . \qquad (7)$$

Comparing the expressions for $\hat{\mathbf{r}}$ and $\boldsymbol{\eta}$ led Rust [11] to suggest the following principle which will guide the selection of estimates for the solution: *An estimate* $\hat{\mathbf{x}}$ *is acceptable only if* $\hat{\mathbf{r}}$ *is a plausible sample from the* $\boldsymbol{\eta}$*-distribution.* It follows then that:

1. The elements of $\hat{\mathbf{r}}$ should be distributed like $n(0,1)$,
2. The elements of $\hat{\mathbf{r}}$ should comprise a white noise time-series,
3. The sum of squared residuals $\hat{\mathbf{r}}^T\hat{\mathbf{r}}$ should lie in some interval $\left[\, m-\kappa\sqrt{2m}\,,\; m+\kappa\sqrt{2m}\,\right]$, with $\mid \kappa \mid \leq 2$.

The last of these conditions follows from the fact that

$$\sum_{i=1}^{m} \eta_i^2 = \boldsymbol{\eta}^T\boldsymbol{\eta} \sim \chi^2(m)\;, \tag{8}$$

whence

$$\mathcal{E}\left\{\boldsymbol{\eta}^T\boldsymbol{\eta}\right\} = m\;,\;\; \sigma^2\left\{\boldsymbol{\eta}^T\boldsymbol{\eta}\right\} = 2m\;. \tag{9}$$

In general κ should be chosen as small as possible without constraining the estimate to violate either of the first two conditions.

The traditional least squares estimate

$$\hat{\mathbf{x}} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}\;\;,\;\; \text{with}\;\;\; \hat{\mathbf{r}}^T\hat{\mathbf{r}} \sim \chi^2(m-n)\;\;, \tag{10}$$

will almost never satisfy the above three conditions because it is almost always the case that $m-n \ll m$. The size of m is limited by the number of measurements that can be made during the course of an experiment and the size of n is chosen as large as possible in order to make the discretized model (2) a good approximation to the original system of integral equations (1).

3 A Test Problem

A test problem capturing many salient features of real instrument correction problems is obtained by discretizing a variant of the well known [10] Phillips equation

$$y(t) = \int_{-3}^{3} K(t,\xi)x(\xi)\,d\xi\;,\;\; -6 \leq t \leq 6\;, \tag{11}$$

where

$$K(t,\xi) = \begin{cases} \frac{1}{6}\left(1+\cos\left[\frac{\pi(\xi-t)}{3}\right]\right)\;, & \mid \xi - t \mid \leq 3 \\ 0\;,\;\; \text{otherwise}\;, \end{cases} \tag{12}$$

and

$$y(t) = \frac{1}{6}\left\{(6-\mid t \mid)\left[1+\frac{1}{2}\cos\left(\frac{\pi t}{3}\right)\right] + \frac{9}{2\pi}\sin\left(\frac{\pi \mid t \mid}{3}\right)\right\}. \tag{13}$$

This problem differs from the original Phillips equation only in the inclusion of the normalizing factor $\frac{1}{6}$ which is needed to assure that the corresponding measuring instrument would not violate physical conservation laws. The solution is

$$x(\xi) = 1+\cos\left(\frac{\pi\xi}{3}\right)\;\;,\;\;\; \mid \xi \mid \leq 3\;. \tag{14}$$

The functions $y(t)$ and $x(\xi)$ are plotted on the right in Figure 1, and on the left are plotted the functions $K(t,\xi_j)$ for $\xi_j = -3.0,\, -1.5,\, 0.0,\, 1.5,\, 3.0$.

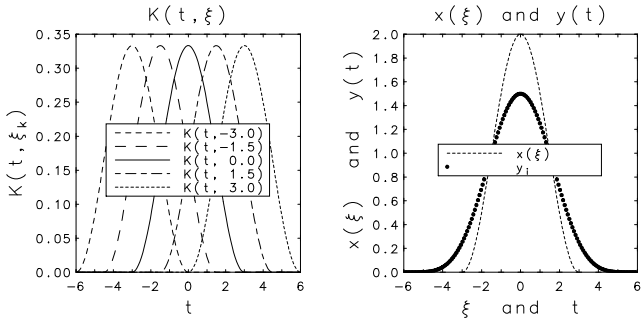


Figure 1: The Phillips Problem

The problem was discretized by choosing $m = 150$ equally spaced t_i on the interval $-5.925 \leq t \leq 5.925$ and using an $n = 121$ point trapezoidal rule on $-3.0 \leq \xi \leq 3.0$ to give

$$\mathbf{y}^* \equiv \mathbf{K}\mathbf{x}^*\;, \tag{15}$$

where $\mathbf{x}^*$ is a 121-vector of $x(\xi_j)$ computed by (14), and $\mathbf{y}^*$ was computed by (15) rather than (13) in order to assure that the ϵ_i were the only errors in the model. The ϵ_i were gotten by random sampling from $N(\mathbf{0},\mathbf{S}^2)$ with

$$\mathbf{S} = \mathbf{diag}(s_1,\; s_2,\; \ldots,\; s_m)\;,\;\; s_i = (10^{-4})y_i^*\;, \tag{16}$$

so the errors in the y_i are in the 4th digit – too small to be discernible in Figure 1. The problem was scaled by the transformations in Eqs. (4) to give a matrix with condition number $\text{cond}(\mathbf{A}) = 2.924 \times 10^9$.

The least squares estimate is shown in Figure 2, where the flattened, dashed curve is $x(\xi)$ and the jagged curve is the estimate $\hat{\mathbf{x}}$. The wild oscillations in the latter are commonly attributed to the ill conditioning of the matrix which greatly magnifies the relatively small errors in the $\mathbf{b}$ vector when the estimate is calculated by (10). But the diagnostics plots given in Figure 3 also demonstrate that the estimate captured variance that properly belongs in the residuals.

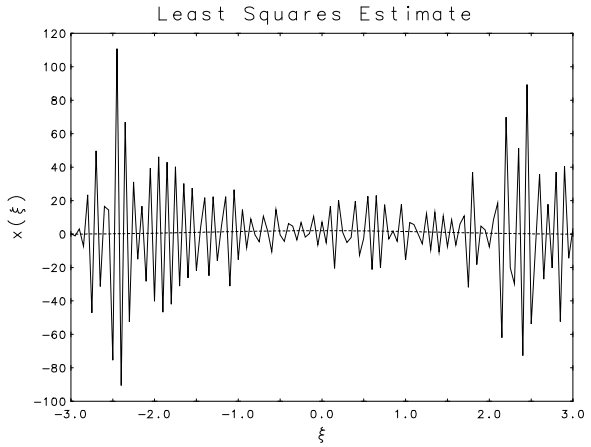


Figure 2: The Linear Least Squares Estimate

4 Estimate Diagnostics

When the true solution is not known, the residual vector provides the only objective guide for assessing the quality of an estimate. The three criteria for acceptable residuals which were given in Section 2 provide quantitative statistical tests for acceptability which are illustrated for the least squares estimate in Figure 3. In the upper left graph, the residuals are plotted as a time-series, i.e., r_i is plotted as a function of the element number i, with sample spacing $\Delta t = \Delta i = 1$. The sum of squared residuals 25.43 is far outside the 2-σ interval

$$\left[\, m - 2\sqrt{2m}\,,\; m + 2\sqrt{2m}\,\right] = [\,115.4\,,\,184.6\,]\,. \quad (17)$$

Inspecting the plot suggests that there are too many small values to have been obtained from a $n(0,1)$ distribution, an impression that is verified by the histogram plot in the upper right graph. The smooth curve in the plot was obtained by renormalizing the $n(0,1)$ distribution with the factor $m = 150$ to make it consistent with the histogram. A formal goodness of fit test gave $\chi^2 = 479.0782$ which is so extremely large that the probablility of a larger value is negligible.

The graph at the lower left of the Figure is a plot of the periodogram of the residual time-series, i.e., a plot of variance versus frequency on the interval $[0,\, 1/2\Delta t]$. An excellent reference on how to estimate and interpret a periodogram is Chapter 7 of Fuller's book on time-series [1]. For the estimate given here the residual time series was zero padded to have 8192 terms and the discrete Fourier transform was calculated at 4096 frequencies $f_k = k/8192$. The periodogram $P(f_k)$ was computed from the squared modulus of the transform. For a white noise time-series, the variance should be distributed evenly over the whole frequency interval. Since the variance is almost completely confined to the first half of the interval, it is clear, even without a formal test, that the least squares residuals do not comprise a white noise series. A formal test confirming this fact is given by the cumulative periodogram which is plotted as a solid curve at the lower right of the Figure.

The cumulative periodogram was computed by

$$C(f_k) = \frac{\sum_{j=1}^{k} P(f_j)}{\sum_{j=1}^{4096} P(f_j)}\;,\quad k = 1, 2, \ldots, 4096\,. \quad (18)$$

The theoretical distribution for white noise would be a line, $C(f) = 2f$. The dashed lines are defined by

$$C_{lo}(f) = -\delta + 2f\;,\quad C^{up}(f) = \delta + 2f\;, \quad (19)$$

where δ is the 5% point for the Kolmogorov-Smirnov statistic for a sample of size $m/2 = 75$. The area between them is a 95% confidence band for white noise. Since the estimate lies outside this band for 64% of the frequencies, the white noise hypothesis is rejected.

5 Regularization

The most widely used method for stabilizing the wildly oscillating least squares estimate of $\mathbf{x}$ is to introduce an *a priori* side constraint of the form

$$\parallel \mathbf{Q}(\mathbf{x} - \mathbf{x}_0) \parallel_2^2 \equiv (\mathbf{x} - \mathbf{x}_0)^T \mathbf{Q}^T \mathbf{Q}(\mathbf{x} - \mathbf{x}_0) \le \beta^2\;, \quad (20)$$

where $\mathbf{x}_0$ is an optional initial estimate of $\mathbf{x}^*$, $\mathbf{Q}$ is a matrix representation of the linear operator for the constraint, and β^2 is a constant determining the strength of the constraint. The estimate for $\mathbf{x}^*$ is obtained by solving

$$\parallel (\mathbf{b} - \mathbf{A}\mathbf{x}) \parallel_2^2 \; + \lambda^2 \parallel \mathbf{Q}(\mathbf{x} - \mathbf{x}_0) \parallel_2^2 \;=\; \min\,, \quad (21)$$

where the parameter λ^2 is a Lagrange multiplier whose value depends on the value of β^2. The solution is

$$\tilde{\mathbf{x}}(\lambda) = \left(\mathbf{A}^T\mathbf{A} + \lambda^2\mathbf{Q}^T\mathbf{Q}\right)^{-1}\left(\mathbf{A}^T\mathbf{b} + \lambda^2\mathbf{Q}^T\mathbf{Q}\mathbf{x}_0\right)\,. \quad (22)$$

The most frequently used choice for $\mathbf{Q}$ is just the n-th order identity matrix. For this choice, the problem can be stated as an augmented linear reqression model

$$\begin{pmatrix} \mathbf{b} \\ \lambda\mathbf{x}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{A} \\ \lambda\mathbf{I}_n \end{pmatrix}\mathbf{x}^* + \begin{pmatrix} \boldsymbol{\eta} \\ \lambda\boldsymbol{\gamma} \end{pmatrix}\,, \quad (23)$$

with

$$\begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\gamma} \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{I}_m & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_n \end{pmatrix}\right]\,, \quad (24)$$

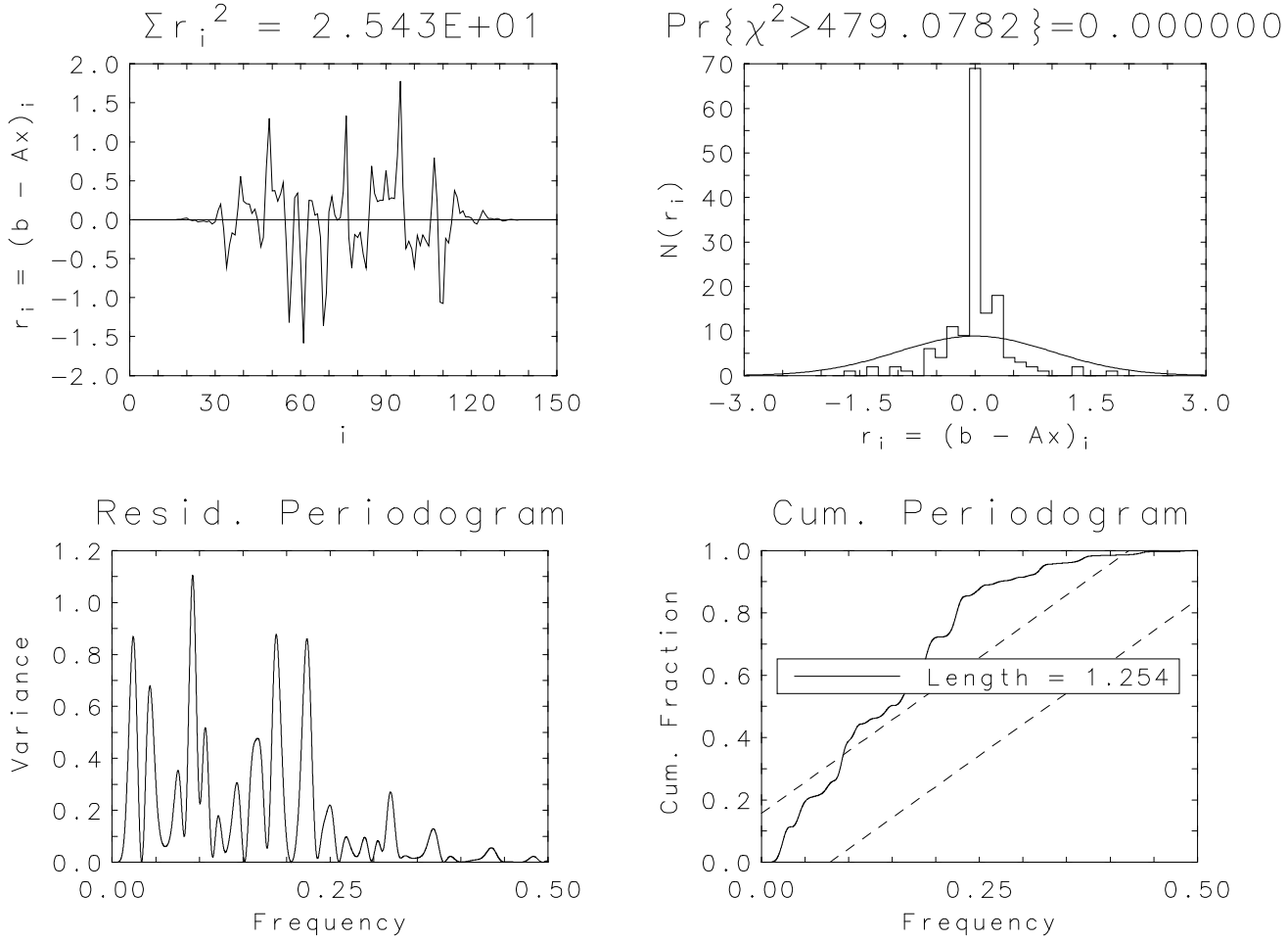


Figure 3: Diagnostics for the Linear Least Squares Estimate

and the parameter λ becomes a weighting constant which must be chosen large enough to damp out wild oscillations in the estimate by keeping it close to $\mathbf{x}_0$, but at the same time small enough to prevent too much growth in the $\| [\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}(\lambda)] \|_2^2$ term. The least squares estimate becomes

$$\tilde{\mathbf{x}}(\lambda) = \left(\mathbf{A}^T\mathbf{A} + \lambda^2\mathbf{I}_n\right)^{-1}\left(\mathbf{A}^T\mathbf{b} + \lambda^2\mathbf{x}_0\right) \ . \qquad (25)$$

In the special case when $\mathbf{x}_0 = \mathbf{0}$ the above procedure is known to numerical analysts as *Tikhonov regularization* and to statisticians as *ridge regression*. This method has been used effectively for almost forty years, but it seems logical that the results would be improved by using an initial estimate better than $\mathbf{x}_0 = \mathbf{0}$. For the test problem given in Section 3, a better initial estimate was gotten by fitting an interpolating spline to the $y(t_i)$ values in the interval $-3 \le t_i \le 3$ and discretizing that spline on $-3 \le \xi_j \le 3$ to get the required 121 elements of $\mathbf{x}_0$. This is the problem that will be discussed in the next 3 sections.

6 The L-Curve Estimate

The success of any regularization calculation depends crucially on the choice of the value for λ. Originally, this was a subjective procedure guided by *a priori* knowledge about the solution, a method still widely used. But determined efforts to automate this choice have been made in both the numerical analysis and the statistics communities. In the former, the currently most widely used method is based on the L-curve first introduced by Lawson and Hanson [9, Chapt. 26] and further developed by Hansen and O'Leary [5, 7]. The L-curve is a plot of $\log_{10} \| \tilde{\mathbf{x}}(\lambda) \|_2$ versus $\log_{10} \| \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}(\lambda) \|_2$ for a range of candidate values for λ. The graph in Figure 4 is an L-curve plot for the test problem. The basic idea of the L-curve method is to choose the λ which constrains the length of the estimate vector $\tilde{\mathbf{x}}(\lambda)$ as much as possible while at the same time increasing the sum of squared residuals as little as possible. Therefore λ is chosen to be the value corresponding to the point in the corner

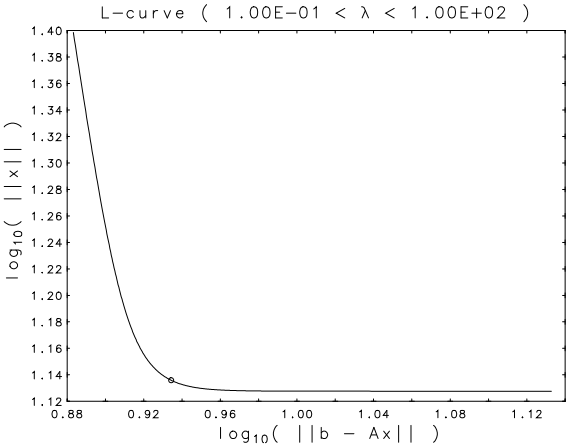


Figure 4: The L-curve for Spline-y Regularization

where the curvature is maximized (the point indicated by the small circle in the plot). For the test problem, this "optimal" value is $\lambda = 0.748$. The corresponding solution is plotted as the solid curve at the upper left of Figure 5 where the true solution is also plotted as a dashed curve. Even if the true solution were not known, it would be clear that a larger value of λ is required to completely suppress the spurious oscillations in the estimate.

The residuals are plotted as a time series at the upper right of the Figure. The sum of squared residuals 73.90 is far below the 2-σ interval (17). The expected value for this quantity, i.e., $m = 150$, gives $\log_{10} \parallel \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}(\lambda) \parallel_2 = 1.0880$ which corresponds to a point far out on the right of the horizontal segment of the L-curve. A histogram plot of the residuals (not shown in the Figure) was similar to the one at the upper right of Figure 3, with far too many values clustered in the central boxes and not enough in the wings. The χ^2 value for testing its goodness of fit to a normal distribution was 147.35 which, with 11 degrees of freedom, would have a negligible probability of occurrence. The periodogram and cumulative periodogram given in the bottom two plots of the Figure indicate consistency with the white noise hypothesis, so the L-curve estimate meets only one of the three criteria given in Section 2 for an acceptable estimate.

7 The Minimum GCV Estimate

The method favored in the statistics community for picking λ is to choose it to minimize the generalized cross-validation (GCV) function

$$G(\lambda) = \frac{\parallel \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}(\lambda) \parallel^2}{\text{trace}\left(\mathbf{I}_m - \mathbf{A}\mathbf{A}^I\right)^2} , \tag{26}$$

where

$$\mathbf{A}^I = \left(\mathbf{A}^T\mathbf{A} + \lambda^2\mathbf{I}_n\right)^{-1}\mathbf{A}^T . \tag{27}$$

The basic idea, first introduced by Wahba [12], is to choose λ to make $\tilde{\mathbf{x}}(\lambda)$ a good predictor for missing data values b_i. More precisely, if $\tilde{\mathbf{x}}^{(k)}(\lambda)$ is the Tikhonov estimate when the kth data point b_k is omitted, then the best value of λ is thought to be the one which minimizes

$$P(\lambda) = \sum_{k=1}^{m} \left[b_k - \left(\mathbf{A}\tilde{\mathbf{x}}^{(k)}(\lambda)\right)_k\right]^2 . \tag{28}$$

It can be shown that the minimizer for $G(\lambda)$ is the same as the minimizer for $P(\lambda)$.

The graph in Figure 6 is a plot of $G(\lambda)$ versus λ for the test problem. The minimum (indicated by the small circle) occurs at $\lambda = 39.250$. The corresponding estimate is plotted as a solid curve in Figure 7 together with the true solution which is plotted as a dashed curve. The two curves are almost indistinguishable, so the minimum GCV estimate is clearly superior to the L-curve estimate, but the correponding residual diagnostics, which are plotted in Figure 8, indicate that it is not an acceptable estimate. The sum of squared residuals 106.7 falls significantly to the left of the acceptable interval (17). The histogram plot of the residuals, shown at the upper right of the Figure, is a very poor approximation to the corresponding normal distribution, and the χ^2 value for comparing the two is 37.0396 which, with 10 degrees of freedom, would occur only 56 times out of a million for samples from the normal distribution. The periodogram and cumulative periodogram for the residual time series indicate consistency with the white noise hypothesis so, like the L-curve estimate, the minimum GCV estimate satisfies only one of the three criteria given in Section 2. Clearly a larger value of λ is needed to further increase the sum of squared residuals and reduce the excessive number of residuals in the central box of the histogram. Inspecting the curve in Figure 6 reveals that increasing λ by a factor of two would not much increase the value of $G(\lambda)$.

8 The "Optimal" Value of λ

Since both the L-curve and the minimum GCV methods gave estimates with unacceptable residual distributions, it was necessary to experiment with several trial values of λ to find one which gave a good estimate. The results

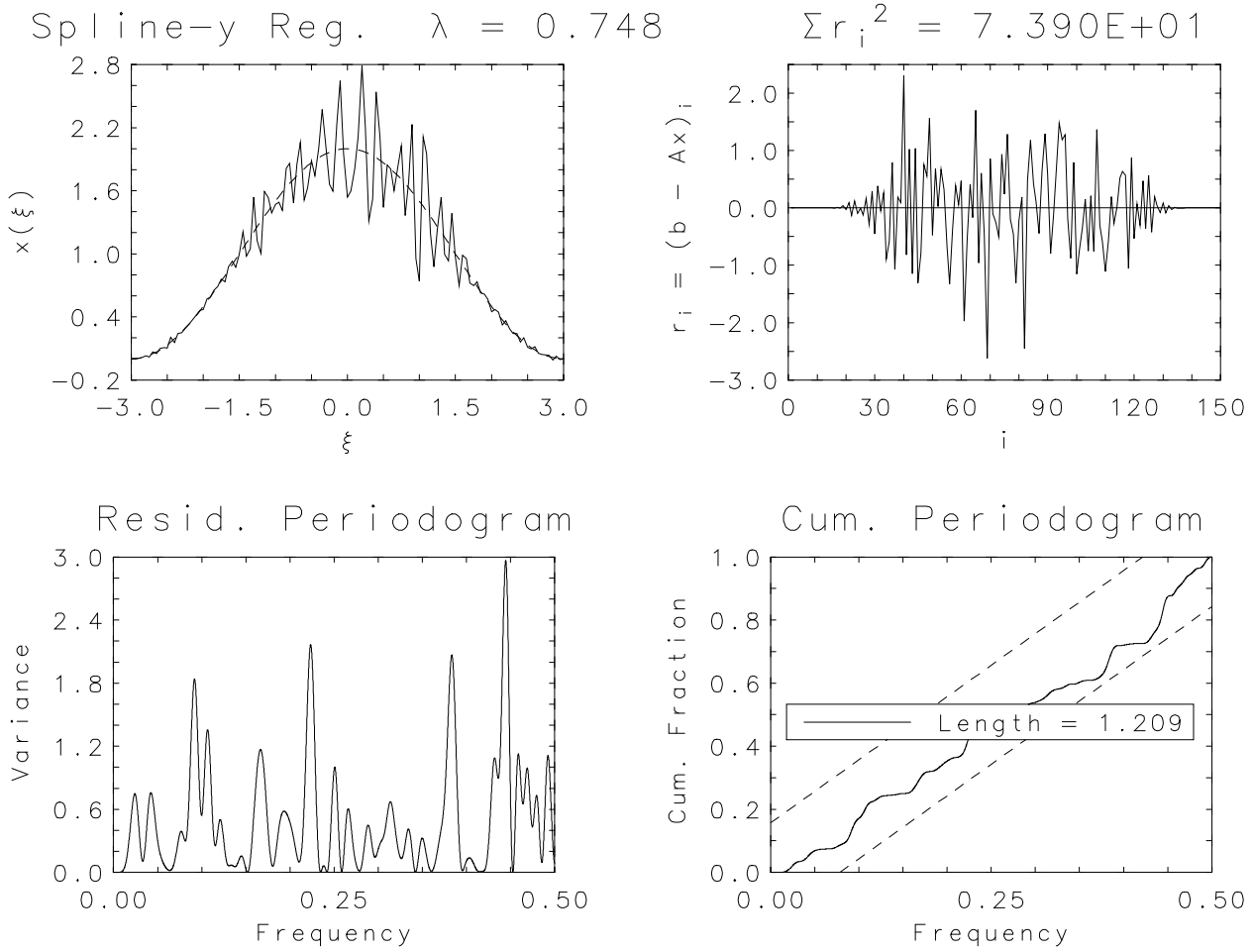


Figure 5: Estimate and Diagnostics for the L-curve Method

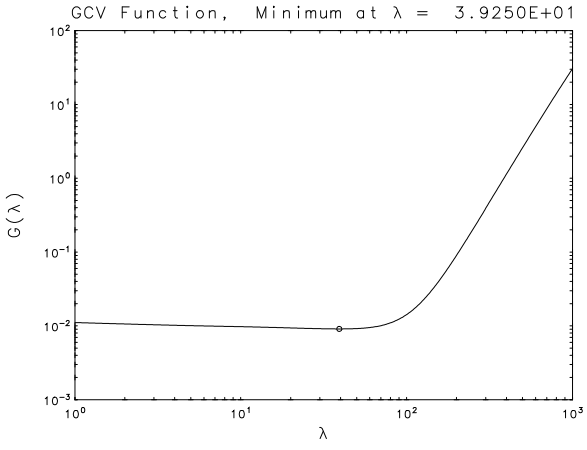


Figure 6: The GCV Function for Spline-y Regularization

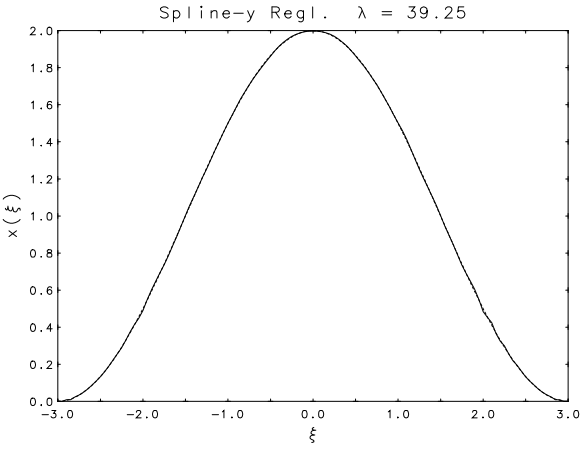


Figure 7: The Minumum GCV Estimate

of that experiment are summarized in Table 1. The 9th column could be included only because the true answer to the problem is known, but it would not be fair to use this knowledge in choosing the best λ. The strategy here

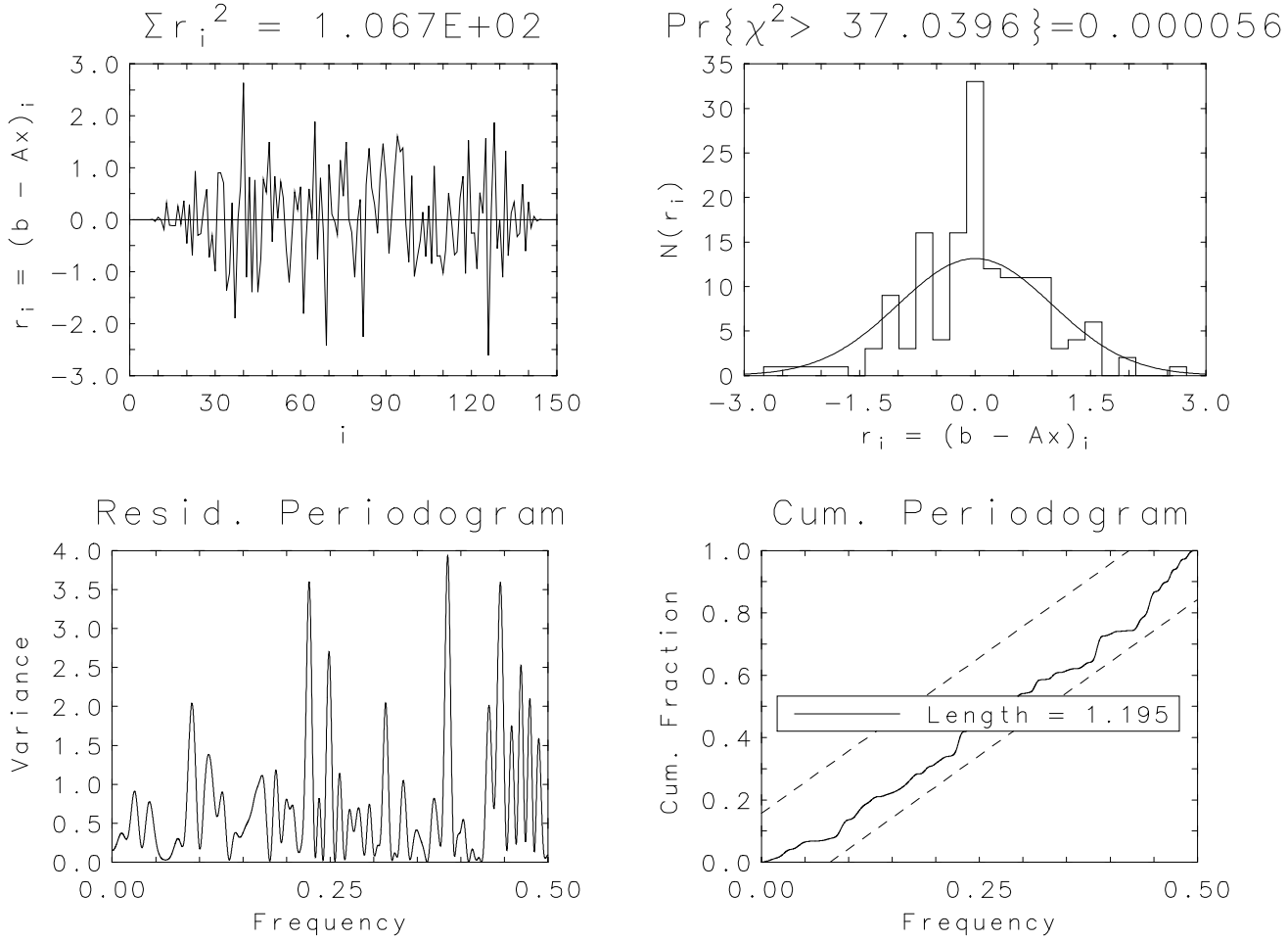


Figure 8: Diagnostics for the Minimum GCV Estimate

was to find the range of values which give residuals that satisfy all three of the criteria in Section 2 and then pick one value from this range as the "optimal" λ.

The first criterion is that the residuals should be distributed like a standard normal distribution. Columns 5, 6 and 7 summarize the results for the χ^2 test of goodness of fit of the residual histogram to the correspodingly renormalized standard normal distribution. For the plotted histograms, the residuals were binned into 25 non-overlapping subintervals of equal length whose union covered $[-r_{\max}, r_{\max}]$, where $r_{\max} = \max_i\{|r_i|\}$, but to do the χ^2 test, subintervals containing fewer than 5 residuals were combined to give a histogram with unequal intervals but with no box containing fewer than 5 counts. Therefore the number of degrees of freedom varied somewhat erratically with changes in the value of λ, so the χ^2 value and its associated probability were not simpe unimodal functions of λ. Accepting 0.05 as the the minimum probability for acceptability means that, except for a small range of values around $\lambda = 85.0$ (row 14), all of the values between $\lambda = 67.5$ (row 7) and $\lambda = 100$ (row 17) gave acceptable residual histograms.

The second criterion in Section 2 is that the residuals should form a white noise time series. The test was based on the cumulative periodogram and the 95% band for white noise whose construction was described in Section 4. Each residual time series was deemed to be white noise if its cumulative periodogram strayed outside the band at no more than 5% of the frequencies, and not to be white noise otherwise. The results of these tests are summarized in Column 8 of the Table which indicates that the second criterion is satisfied only for values of λ in the interval $[\,0.748\,,\ 80.0\,]$. Actually, 0.748 was the smallest non-zero value that was tried, so the acceptable range might extend to somewhat smaller values, but these would fail to satisfy either of the other two criteria. In fact, the intersection of this interval with the acceptable interval from the preceding paragraph, i.e., $[\,67.5\,,\ 100.0\,]$, gives $67.5 \le \lambda \le 80.0$ as the range of values which satisfy both of the first two criteria.

Table 1: Diagnostics for the Spline-y Regularization Experiment

1	2	3	4	5	6	7	8	9	10
	Method	λ	$\sum r_i^2$	χ^2	ndf	*Prob*	wn	$\|\Delta x\|_{rms}$	Fig.
1	Lst. Sqr.	0.000	25.4	479.08	8	0.0000	no	30.24400	3
2	L-curve	0.748	73.9	147.35	11	0.0000	**yes**	0.23961	5
3	Min. GCV	39.250	106.7	37.04	10	0.0001	**yes**	0.00374	8
4	Trial λ	55.000	113.6	30.16	13	0.0045	**yes**	0.00268	
5	Trial λ	60.000	**117.0**	31.92	13	0.0025	**yes**	0.00246	
6	Trial λ	65.000	**121.1**	24.89	13	0.0239	**yes**	0.00228	
7	Trial λ	**67.500**	**123.5**	20.87	13	**0.0755**	**yes**	0.00220	
8	Trial λ	**70.000**	**126.1**	19.15	14	**0.1591**	**yes**	0.00213	
9	Trial λ	**72.500**	**129.0**	15.36	14	**0.3542**	**yes**	0.00206	
10	Trial λ	**75.000**	**132.2**	15.75	14	**0.3289**	**yes**	0.00200	
11	Trial λ	**77.500**	**135.6**	17.56	15	**0.2867**	**yes**	0.00195	9
12	Trial λ	**80.000**	**139.5**	16.31	14	**0.2948**	**yes**	0.00190	
13	Trial λ	82.500	**143.6**	18.18	15	**0.2533**	no	0.00186	
14	Trial λ	85.000	**148.2**	26.34	15	0.0346	no	0.00181	
15	Trial λ	87.500	**153.1**	25.11	16	**0.0679**	no	0.00178	
16	Trial λ	90.000	**158.4**	21.42	16	**0.1631**	no	0.00175	
17	Trial λ	100.000	**184.4**	26.24	16	**0.0508**	no	0.00167	
18	Trial λ	110.000	219.0	46.10	14	0.0000	no	0.00164	
19	Trial λ	120.000	264.0	88.72	14	0.0000	no	0.00166	
20	Trial λ	130.000	321.2	207.98	15	0.0000	no	0.00172	

Column 1 contains integers indexing the different trials.

Column 2 gives the method used for choosing the value of λ.

Column 3 gives the value of λ. Boldface values give residuals satisfying all three criteria in Section 2.

Column 4 gives the sum of squared residuals. Boldface values give residuals which satisy criterion 3 in Section 2.

Column 5 gives the χ^2 value for testing the goodness of fit of the residual histogram to the corresponding normal distribution.

Column 6 gives the number of degrees of freedom used in the goodness of fit test.

Column 7 gives the probability of obtaining a χ^2 value as high as or higher than the one in column 5. Boldface values give residuals which satisfy criterion 1 in Section 2.

Column 8 tells whether or not the cumulative periodogram of the residual time series indicated consistency with the white noise hypothesis. Boldface values indicate consistency with criterion 2 in Section 2.

Column 9 gives the root mean square average magnitude of the errors in the elements of the estimate, i.e.,

$$|\Delta x|_{rms} = \sqrt{\frac{1}{n}\sum_{j=1}^{n} |\tilde{x}_j(\lambda) - x_j^*|^2}\ , \tag{29}$$

where $\tilde{x}_j(\lambda)$ is the jth element of $\tilde{\mathbf{x}}(\lambda)$ and x_j^* is the jth element of the true solution $\mathbf{x}^*$.

Column 10 gives the Figure number for the corresponding diagnostic plots.

The third criterion in Section 2 restricts the sum of squared residuals to lie inside the 2-σ interval for its expected value, i.e., $\left[\, m - \kappa\sqrt{2m}\,,\ m + \kappa\sqrt{2m}\,\right]$, with $|\ \kappa\ | \leq 2$. The sums of squared residuals are tabulated in Column 4 of the Table. Choosing $\kappa = 2$ gives (17) as the acceptable interval for the sum of squared residuals, and comparing this range with the values in Column 4 reveals that all λ in the interval $[\,60.0\,,\ 100.0\,]$ produce acceptable values. This interval completely contains the one given in the preceding paragraph for satisfying both of the first two criteria, so any λ value in the interval $[\,67.5\,,\ 80.0]$ will give residuals which satisfy all three of the criteria given in Section 2.

The acceptable interval for λ can be further reduced by strengthening one or more of the constraints imposed by the three criteria. For example, requiring that the probability value in Column 7 be greater than 0.10 restricts the acceptable interval to $[\,70.0\,,\ 80.0\,]$. Also, choosing $\kappa = 1$ in the third criterion constrains the sum of squared residuals to lie in the interval

$$\left[\, m - \sqrt{2m}\,,\ m + \sqrt{2m}\,\right] = [\,132.7\,,\,167.3\,]\,, \qquad (30)$$

which in turn requires $75.0 \leq \lambda \leq 90.0$. Intersecting the latter interval with the former resticts λ to the range $75.0 \leq \lambda \leq 80.0$. This is a fairly narrow interval, and there is no compelling reason to choose any one particular value in it over any other, so the midpoint $\lambda = 77.5$ was taken to be the "optimal" value. The corresponding estimate is graphically indistinguishable from the true solution. The diagnostic plots are given in Figure 9.

It will be noted that the "optimal" λ does not minimize the root mean square error (29) in the estimate. The values in column 9 of the Table indicate that this minimum is approximately obtained when $\lambda = 110.0$. But this estimate does not satisfy any of the three criteria in Section 2. While this is disturbing, one must remember that for a real problem one could not compute this rms error. And for the "optimal" estimate this error is acceptable even if it is not minimal.

9 Truncating the Singular Value Decomposition

Another method for stabilizing the solution to (5) is to truncate the *singular value decomposition* (SVD). The matrix $\mathbf{A}$ has a unique factorization

$$\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{\Sigma} \\ \mathbf{0} \end{pmatrix}\mathbf{V}^T\,, \quad \mathbf{\Sigma} = \mathbf{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)\,, \quad (31)$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ are the singular values of $\mathbf{A}$, and

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}_m = \mathbf{U}\mathbf{U}^T\,, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}_n = \mathbf{V}\mathbf{V}^T\,. \qquad (32)$$

Substituting (31) into (5) and premultiplying by $\mathbf{U}^T$ gives

$$\mathbf{U}^T\mathbf{b} = \begin{pmatrix} \mathbf{\Sigma} \\ \mathbf{O} \end{pmatrix}\mathbf{V}^T\mathbf{x}^* + \mathbf{U}^T\boldsymbol{\eta}\,, \quad \mathbf{U}^T\boldsymbol{\eta} \sim N(\,\mathbf{0}\,,\,\mathbf{I}_m\,)\,, \qquad (33)$$

with the distribution of the $\mathbf{U}^T\boldsymbol{\eta}$ vectors unchanged because premultiplication by an orthogonal matrix simply rotates all the vectors in the distribution through the same angle.

If $\mathbf{A}$ has full rank, then it is easy to see that the least squares estimate (10) can be written

$$\hat{\mathbf{x}} = \mathbf{V}\left(\,\mathbf{\Sigma}^\dagger\,,\,\mathbf{0}\,\right)\mathbf{U}^T\mathbf{b}\ \,, \qquad (34)$$

where

$$\mathbf{\Sigma}^\dagger = \mathbf{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_n}\right)\,. \qquad (35)$$

It is instructive to rewrite these two equations in the form

$$\left(\mathbf{V}^T\hat{\mathbf{x}}\right)_i = \frac{\left(\mathbf{U}^T\mathbf{b}\right)_i}{\sigma_i}\,, \quad i = 1, 2, \cdots, n\,, \qquad (36)$$

and to note that the minimum sum of squared residuals can be written

$$r^2_{\min} = \|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}\|^2 = \sum_{i=n+1}^{m}\left(\mathbf{U}^T\mathbf{b}\right)_i^2\,. \qquad (37)$$

Since the least squares estimate contains spurious large oscillations, and since its sum of squared residuals is unacceptably small, it appears that Eqs. (36) capture some of the variance that properly belongs in the sum (37).

If $\mathbf{A}$ has less than full rank, then a minimum length least squares estimate $\tilde{\mathbf{x}}$ can be computed by using Eq. (34) with

$$\mathbf{\Sigma}^\dagger = \mathbf{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_p}, 0 \ldots, 0\right)\,, \qquad (38)$$

where $p = rank(\mathbf{A}) < n$. In this case Eqs. (36) are replaced by

$$\left(\mathbf{V}^T\tilde{\mathbf{x}}\right)_i = \begin{cases} \dfrac{\left(\mathbf{U}^T\mathbf{b}\right)_i}{\sigma_i} & ,\quad i = 1, 2, \cdots, p\,, \\ 0 & ,\quad i = p+1, \cdots, n\,, \end{cases} \qquad (39)$$

and Eq. (37) is replaced by

$$\tilde{r}^2 = \|\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}\|^2 = \sum_{i=p+1}^{m}\left(\mathbf{U}^T\mathbf{b}\right)_i^2\,. \qquad (40)$$

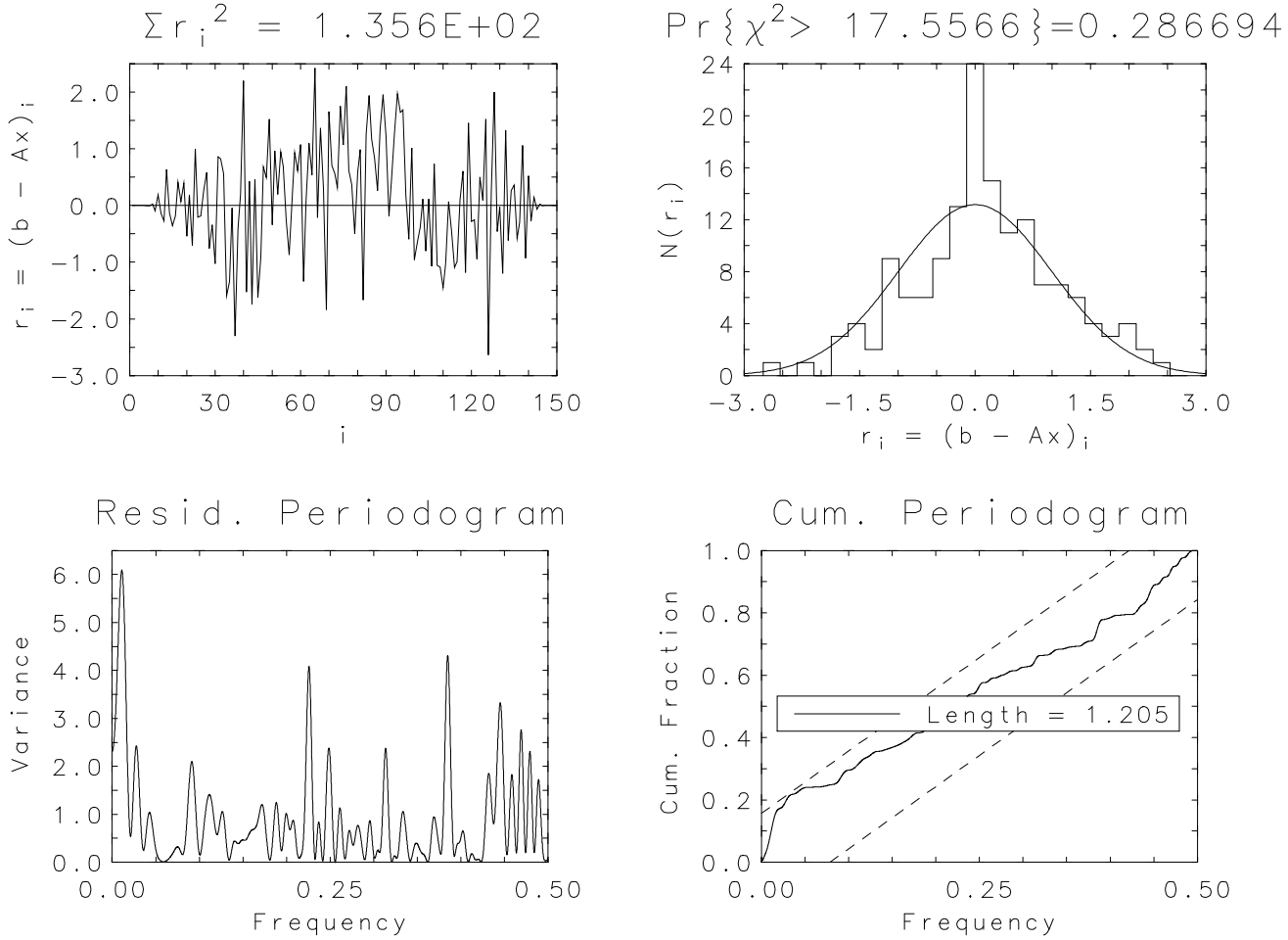


Figure 9: Diagnostic Plots for Spline-y Regularization with $\lambda = 77.5$

Since real computer calculations never give exact zero singular values, and since the ill-conditioning gives $\sigma_1 \gg \sigma_n$, a natural hypothesize was that the wild oscillations in the least squares estimate are caused by inverting small, inaccurately determined σ_j whose values should properly be zero. If so, the estimate could be stabilized by setting $\sigma_j = 0$ for all singular values below some threshold σ_p and computing $\tilde{\mathbf{x}}$ from Eqs. (39) rather than Eqs. (36).

The above idea was first suggested by Golub and Kahan [2], and further developed by Hanson [8]. The conventional view held that $\mathbf{A}$ was rank deficient, and that the truncation should determine its "numerical rank." Most methods (e.g., [3]) attempted to find a clear gap in the singular value distribution and to zero all those on the low side. Many years elapsed before it was widely appreciated [4] that, even though truncation stabilizes the estimate, the matrices for real-world problems almost never display such a gap. The dilemma is illustrated by Figure 10 where the singular values for the test problem are plotted as discrete squares. There is no gap where the values plunge precipitously. The largest and smallest are $\sigma_1(\mathbf{A}) = 3.39 \times 10^7$ and $\sigma_{121}(\mathbf{A}) = 1.16 \times 10^{-2}$ which give $cond(\mathbf{A}) = 2.92 \times 10^9$, but the relative accuracy of the calculation was $\epsilon_{mach} = 2.22 \times 10^{-16}$, so σ_{121} is 7 orders of magnitude greater than the effective zero level. Thus, there is no reason to assume that $rank(\mathbf{A}) < n$, but some truncation is needed to prevent the estimate from capturing variance that belongs in the residuals. Fortunately, this result can be obtained by leaving the singular values unchanged and truncating the rotated right hand side vector $\mathbf{U}^T\mathbf{b}$.

10 Truncating the Vector $\mathbf{U}^T\mathbf{b}$

Figure 10 also shows the first n elements of the vector $|\mathbf{U}^T\mathbf{b}|$ plotted as small circles connected by straight line segments. There is a sharp break in the distribution at $j = 36$. Before the break, the upper envelope is

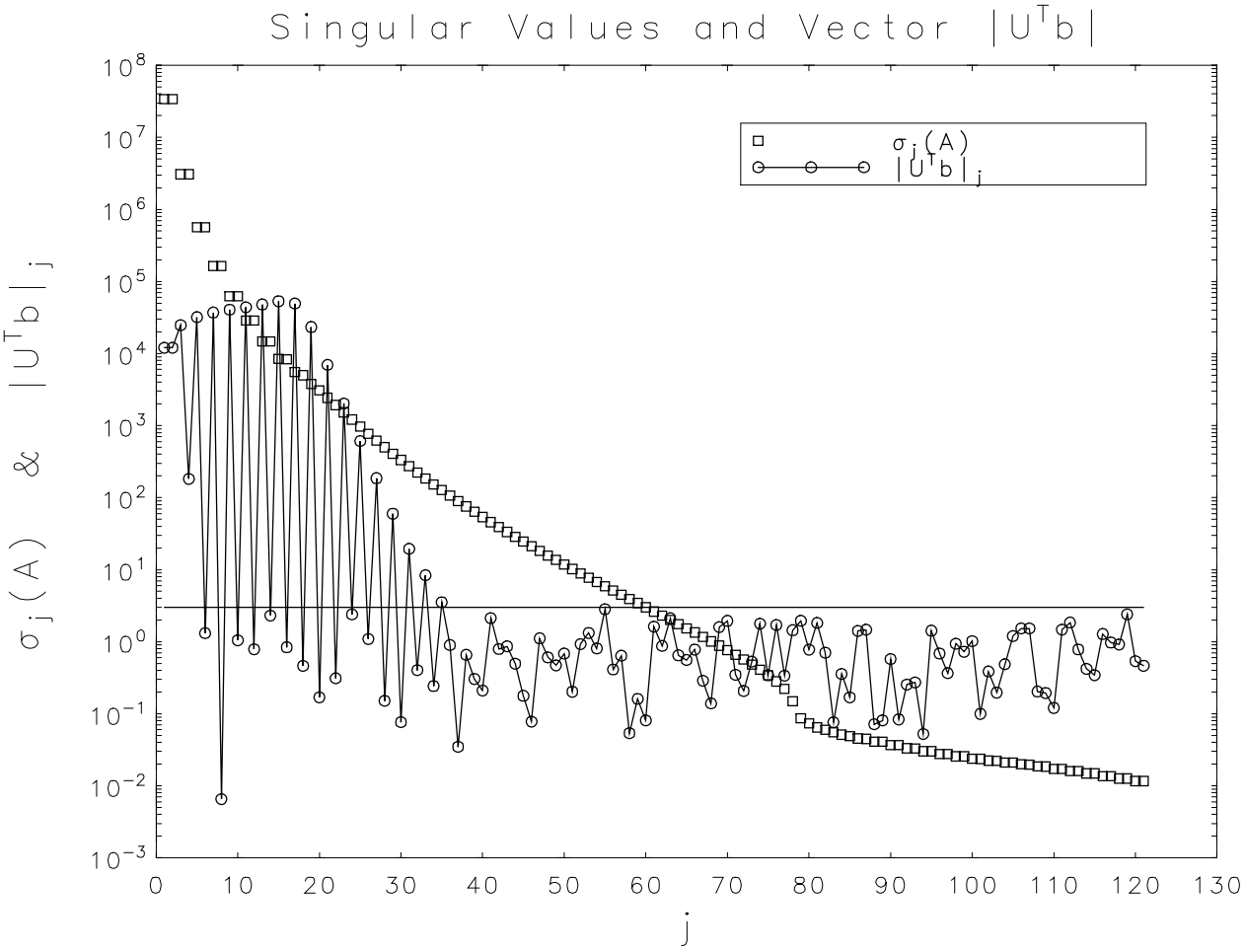


Figure 10: Singular Values and First n Elements of Vector $|\mathbf{U}^T\mathbf{b}|$ for the Test Problem

decreasing more rapidly than the corresponding singular values. After the break, the distribution is flat with all values below the line $|\mathbf{U}^T\mathbf{b}| = 3$. That line is, by Equations (33), just the 3-σ level for the magnitudes of the errors $(\mathbf{U}^T\boldsymbol{\eta})_i$. Thus, the flat segment contains elements of $\mathbf{U}^T\mathbf{b}$ that are dominated by measuring errors. In 1998 Rust [11] suggested a criterion for truncating the SVD by altering the transformed measurements, which contain random errors, rather than the matrix, which is assumed to be exact. *Rather than zeroing some of the singular values, one should instead zero those elements of* $\mathbf{U}^T\mathbf{b}$ *that are dominated by the random errors.* The idea is to pick a truncation level τ and require $\tilde{\mathbf{x}}(\tau)$ to satisfy

$$\left(\mathbf{V}^T\tilde{\mathbf{x}}(\tau)\right)_i = \begin{cases} \dfrac{\left(\mathbf{U}^T\mathbf{b}\right)_i}{\sigma_i} & , \quad \text{if } |\mathbf{U}^T\mathbf{b}|_i > \tau \,, \\ & \quad\; i = 1, 2, \ldots, n \,, \\ 0 & , \quad \text{otherwise} \,. \end{cases} \tag{41}$$

The sum of squared residuals is then given by

$$\tilde{r}^2 = \|\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}(\tau)\|^2 = \sum_{i \,\in\, \mathcal{I}} \left(\mathbf{U}^T\mathbf{b}\right)_i^2 + \sum_{i=n+1}^{m} \left(\mathbf{U}^T\mathbf{b}\right)_i^2 \,, \tag{42}$$

where the indexing set for the first sum is

$$\mathcal{I} = \left\{ \, i \;\middle|\; |\mathbf{U}^T\mathbf{b}|_i \le \tau \,, \, i = 1, 2, \ldots, n \, \right\} \,. \tag{43}$$

The success of the proposed method depends crucially on the choice of the truncation level τ. A safe and effective approach is to try several values, and use the criteria in Section 2 to make the final choice. A guideline for choosing a lower bound for the value of τ is the fact that most experimentalists would be reluctant to claim that a measured value is significantly different from zero if its magnitude does not exceed 3 standard deviations for the error in the measurement. Since each $|\mathbf{U}^T\mathbf{b}|_i$ is scaled in units of one standard deviation of its own random error, it suffices to choose $\tau = 3.0$. Of course, to be

Table 2: Diagnostics for the Truncated $\mathbf{U}^T\mathbf{b}$ and Truncated SVD Methods

1	2	3	4	5	6	7	8	9	10
	Method	τ / p	$\sum r_i^2$	χ^2	ndf	*Prob*	wn	$\lvert\Delta x\rvert_{rms}$	Fig.
1	Lst. Sqr.	$\tau =$ 0.0	25.4	479.08	8	0.0000	no	30.24400	3
2	Trial τ	$\tau =$ 2.3	104.6	17.87	12	**0.1196**	no	17.23000	
3	Trial τ	$\tau =$ **2.5**	**121.3**	12.77	13	**0.4661**	**yes**	.04360	11
4	Trial τ	$\tau =$ **3.0**	**129.3**	13.88	13	**0.3823**	**yes**	.00153	12
5	Trial τ	$\tau =$ **4.0**	**141.8**	16.19	12	**0.1826**	**yes**	.00220	
6	Trial τ	$\tau =$ 9.0	211.4	17.27	12	**0.1398**	**yes**	.00406	
7	Trial τ	$\tau = 20.0$	589.1	11531	16	0.0000	no	.00797	
8	L-curve	$p = 70$	75.0	181.90	10	0.0000	**yes**	.32000	
9	Min. GCV	$p = 35$	112.4	17.67	14	**0.2224**	**yes**	.00156	

All Columns the same as in Table 1 except Column 3 which here contains the truncation parameter.

Rows 1-7 are for the truncated $\mathbf{U}^T\mathbf{b}$ method. Rows 8-9 are for the conventional truncated SVD method.

safe, it might be wise to also try even smaller values like $\tau = 2.5$. The choice for an upper bound for τ might similarly be guided by the fact that most experimentalists would readily belive that a measured value greater than the 6-σ level is statistically significant. Since there are only n discrete $|\mathbf{U}^T\mathbf{b}|_i$, there are only n possible truncated estimates $\tilde{\mathbf{x}}(\tau)$, and, in genrerall, only a few of them will correspond to truncation levels between $\tau = 3.0$ and $\tau = 6.0$.

Table 2 gives, for the test problem, the diagnostics for all possible truncated estimates with $2.3 \le \tau \le 20.0$. The two extreme values and the least squares ($\tau = 0.0$) estimate are included only for the sake of comparisons. The judicious application of a straight edge and pencil to Figure 10 reveals that the difference between the $\tau = 20.0$ and the $\tau = 9.0$ estimates was caused by the exclusion from the former of the single element $(\mathbf{U}^T\mathbf{b})_{31} = 19.4342$. The $\tau = 9.0$ estimate, which satisfies only the first two of the three criteria in Section 2, differs from the $\tau = 4.0$ estimate because it also excludes the element $(\mathbf{U}^T\mathbf{b})_{33} = 8.34205$. The $\tau = 4.0$ estimate satisfies all of the three criteria as do the estimates for $\tau = 3.0$ and $\tau = 2.5$. The $\tau = 4.0$ and $\tau = 3.0$ estimates differ because the latter includes one additional element $(\mathbf{U}^T\mathbf{b})_{35} = 3.53193$. The $\tau = 2.5$ estimate includes one other additional element $(\mathbf{U}^T\mathbf{b})_{55} = -2.82565$, and the $\tau = 2.3$ contains still one further additional element $(\mathbf{U}^T\mathbf{b})_{119} = 2.39588$. The latter estimate satisfies only the first of the three criteria.

Only the $\tau = 2.5$, 3.0, and 4.0 estimates satisfy all three of the criteria in Section 2. The $\tau = 2.5$ estimate and the true solution are plotted in Figure 11. Some of the spurious oscillations persist, so this estimate can be

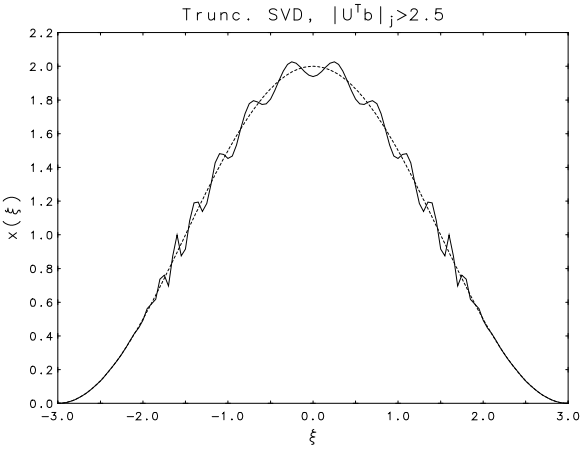


Figure 11: Truncated $\mathbf{U}^T\mathbf{b}$ Estimate for $\tau = 2.5$

rejected. This assumes an *a priori* knowledge that the solution should not contain oscillations, but such judgements are common in solving ill-posed problems. The $\tau = 3.0$ and $\tau = 4.0$ estimates are both graphically indistinguishable from the true solution, and the numerical diagnostics indicate that either would be acceptable. The $\tau = 4.0$ estimate gives a better sum of squared residuals, but the $\tau = 3.0$ estimate gives a markedly better χ^2 probability. It is probably best to choose the alternative that makes the smallest change in the given data, so $\tau = 3.0$ was taken to be the optimal value. The plotted diagnotics are given in Figure 12. Note that $\tau = 3.0$ minimizes $|\Delta x|_{rms}$ whose value was slightly better than the best obtained by spline-y regularization (cf., Table 1, row 18).

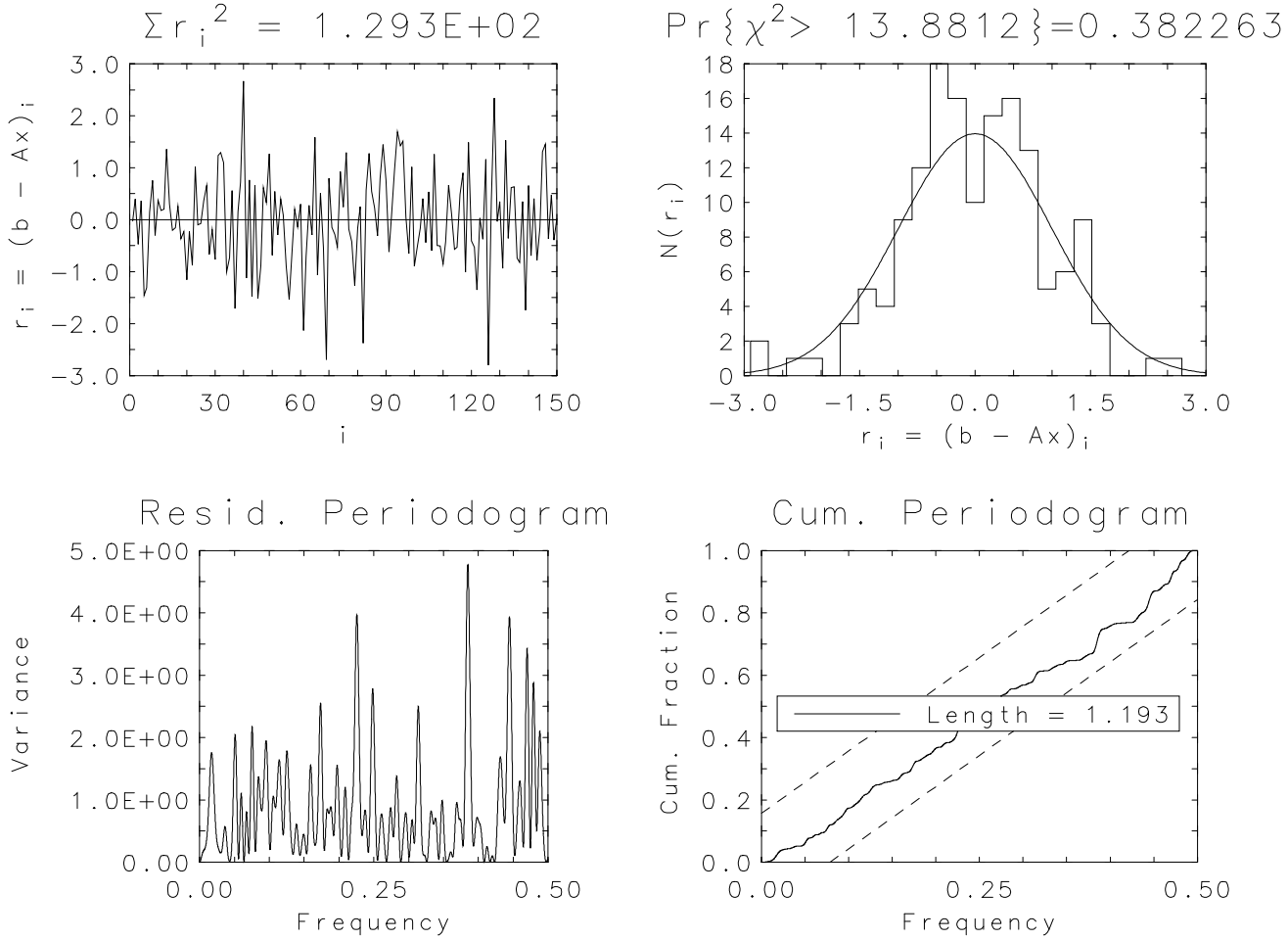


Figure 12: Diagnostic Plots for the Truncated $\mathbf{U}^T\mathbf{b}$ Estimate with $\tau = 3.0$

11 Comparison with the Standard SVD Truncation

The conventional method for truncating the SVD effectively alters the matrix, which is assumed to be known exactly, or at least to greater accuracy than the data vector, in order to accommodate the errors in the latter. Thus, the same matrix may be assigned different "numerical ranks" for different measurement vectors. Nevertheless, the method can give good results and has enjoyed much success even though some of the strategies for picking the truncation order p have been somewhat muddled.

The two strategies given in Sections 6 and 7 have both been adapted [6] to pick the order p. It is straightforward to pick the minimal GCV value, by replacing Eqn. (26) with

$$G(p) = \frac{\| \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}(p) \|^2}{\text{trace}\,(\mathbf{I}_m - \mathbf{A}\mathbf{A}^I)^2} \ , \quad p = 1, 2, \ldots, n \ , \qquad (44)$$

where $\tilde{\mathbf{x}}(p)$ is the estimate computed from Eqs. (39), and evaluating $G(1), G(2), \ldots, G(n)$. The L-curve method is somewhat trickier, because the plot of $\log_{10}(\|\tilde{\mathbf{x}}(p)\|)$ versus $\log_{10}(\|\mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}(p)\|)$ contains n discrete points in a L-shaped pattern. To find the maximum curvature, it is necessary to fit a smoothing spline to those points and determine where the curvature of that spline is maximized. It is then necessary to choose which of those points is closest to the point on the spline where the curvature is maximized.

The numerical diagnostics for the L-curve and minimum GCV estimates are given in rows 8 and 9 of Table 2. The L-curve method gives $p = 70$ which corresponds to an element far out in the noise dominated, flat segment of the $|\mathbf{U}^T\mathbf{b}|$ plot in Figure 10. The consequence of including the noise dominated elements to the left of this point is an oscillatory estimate similar to the one plotted in Figure 5 and residuals which satisfy only the second of the three criterion in Section 2. The minimum GCV method does much better, giving $p = 35$, which is just

before the onset of the flat segment of the $|\mathbf{U}^T\mathbf{b}|$ plot. The set of elements used in the estimate is similar to the one used in the $\tau = 3.0$ truncated $|\mathbf{U}^T\mathbf{b}|$ estimate. Both estimates reject all of the elements with $j > 35$. They differ only in that the truncated SVD method retains all of the elements with $j \leq 35$, but the truncated $|\mathbf{U}^T\mathbf{b}|$ method rejects the subset of 15 elements which also satisfy $|\mathbf{U}^T\mathbf{b}|_j \leq 3.0$. Since their corresponding singular values are about 3 orders of magnitude greater than the $|\mathbf{U}^T\mathbf{b}|_j$, the inclusion of the extra 15 ratios in Eqs. (39) does not change the truncated SVD estimate very much. Thus its $|\Delta x|_{rms}$ is almost as good as the one for the $\tau = 3.0$ estimate. But, its sum of squared residual falls just below the acceptable interval (17), so even though the residuals satisfy the other two criteria from Section 2, the estimate would, in the absence of a knowledge of the true solution, have to be rejected.

12 Discussion and Conclusions

The ill-conditioning of the linear regression model obtained by discretizing an ill-posed problem produces, in the least squares estimate, a huge magnification of the measurement errors. Minimizing the sum of squared residuals allows the estimate to capture variance that properly belongs in the residuals. Therefore it is necessary to seek biased estimates which either append side constraints (regularization) or which discard some components of the variance in the measurements (truncating the SVD). In the first case it is necessary to choose a free parameter which determines the relative weightings of the regression equations and the side constraints, and in the second case it is necessary to choose a cut-off parameter which determines how much of the measurement variance to discard. The guiding principle adopted here for making such choices is that *a good estimate should produce a residual vector that resembles a sample from the measurement error distribution.*

For normally distributed measurement errors, with known variance matrix, the regression model can be rescaled to give errors that are independently distributed with the standard normal distribution. The above guiding principle can then be used to specify three statistical criteria which should be satisfied by the residuals for an acceptable estimate. These criteria were quantified and applied to estimates obtained by both methods for a test problem designed to resemble a real instrument correction problem. They were used both to test previously suggested strategies for picking the free parameters and to choose optimal values for them.

The regulatization method used here applied side constraints to bias the solution toward an initial estimate obtained by fitting a spline to the measurement vector. Two widely used strategies for choosing the regularization parameter gave residuals satisfying only one of the three statistical criteria. The L-curve method failed because it kept the sum of squared residuals too small while minimizing the length of the estimate vector. The tight constraint on the length of the residual vector must be relaxed so that it can capture the variance that causes the wild oscillations in the estimate. The minimum GCV estimate was better than the L-curve estimate, but it did not produce acceptable residuals either. The problem seems to be that the minimum of the GCV function occurred in a relatively flat segment of the curve that allows large deviations from the minimum point without changing the cross validation index very much.

An acceptable regularized solution was found by calculating estimates for a range of parameter values, all larger than the L-curve and minimum GCV values. Each of the three statistical criteria was satisfied in a different subinterval of the range, but the intersection of the three subintervals defined a narrow range which gave residuals satisfying all three criteria. The optimal value was chosen to be the one that gave the residual histogram best fitting the standard normal distribution. The corresponding estimate gave a good root mean square average error, but the smallest value of that quantity was obtained at a parameter value on the high side of all three subintervals. This behavior has been observed for two other test problems, so it may be a universal property of regularization methods. If so, it may be possible to find some indicator for the parameter value which isolates this minimum error, but until such an indicator is found, prudence recommends the value which gives residuals best satisfying the three statistical criteria.

Traditionally the SVD was truncated by zeroing the smaller singular values and transferring the corresponding components of $\mathbf{U}^T\mathbf{b}$ from the estimate to the residuals. Early ideas associating the truncation with the "numerical rank" of the matrix could not be sustained because the matrix is seldom rank deficient. Since the instability of the estimate is caused by errors in the right hand side vector rather than from any deficiencies in the matrix, it is more logical to truncate the former rather than the latter. Scaling the model to give errors that are independently normally distributed and substituting the SVD for the scaled matrix rotates the problem into a coordinate system in which it is easy, using reasoning familiar to any experimental scientist, to identify the components of $\mathbf{U}^T\mathbf{b}$ that are completely dominated by measurement errors. There are only n possible truncations, and only a few of these fall in the acceptable range $2.5 \leq \tau \leq 6.0$. Ambiguities in the choice of τ can

be resolved by using the three statistical criteria for acceptable residuals. When this approach was applied to the test problem, only two possible truncations were acceptable. The optimal τ was taken to be the one whose residual distribution gave the best fit to the appropriately rescaled standard normal distribution. This estimate was also the one which minimized the root mean square average error.

Although "numerical rank" did not prove to be useful for truncating the SVD, the order truncation approach has some merit. Both the L-curve and minimum GCV strategies have been adapted to choose the truncation order p [6], but for the test problem, both methods gave residuals which satisfied only one of the three statistical criteria. The L-curve estimate failed to suppress all of the spurious oscillations because it did not sufficiently relax the constraint on the length of the residual vector. The minimum GCV estimate was quite good even though its residuals failed to satisfy two of the three criteria. Its root mean square average error was almost as small as the one for the optimal τ-truncation, mainly because all of its discarded elements were also discarded by the latter method. Unfortunately, it retained a few elements discarded by the latter which did not change the estimate very much but did render the residuals statistically unacceptable. It might be possible to use the GCV function for automatically choosing the truncation parameter by considering (44) as a function of τ rather than p.

There are circumstances in which it is necessary to use order truncation in conjunction with τ-truncation. For the test problem, the flat segment of the $\mathbf{U}^T\mathbf{b}$ plot contained only 86 elements, all of them smaller in magnitude than 3.0. For larger problems, with longer flat segments, it becomes increasingly probable that normally distributed random errors will produce larger elements in this noise saturated segment of the plot. Choosing τ large enough to suppress such rare values may also discard some earlier elements that are not mostly noise. Therefore, to keep the value of τ reasonable, it is necessary to impose an additional order truncation in the flat segment. But this approach could also prove disasterous if the true solution has highly oscillatory components corresponding to higher order singular vectors. In such cases, the τ-truncation alone might be vital for isolating those components.

Finally it should be noted that even though τ-truncation can stabilize the estimate, it cannot guarantee that it will be close to the true solution because the noise in the measurements might be large enough to overwhelm important components of the signal being measured. The only way to retrieve such components would be to repeat the measurements with more precision.

Acknowledgements

The author would like to thank Drs. Mark S. Levenson, Anastase Nakassis, and Dianne P. O'Leary for useful advice and criticism.

References

[1] Fuller, W. A. (1976) *Introduction to Statistical Time Series*, John Wiley & Sons, New York.

[2] Golub, G. and Kahan, W. (1965) *J. SIAM Numer. Anal. Ser. B*, **2**, pp. 205-224.

[3] Golub, G., Klema, V., and Stewart, G. W. (1976) *Rank Degeneracy and Least Squares Problems*, Technical Report TR-456, Department of Computer Science, University of Maryland.

[4] Hansen, P. C. (1987) *BIT*, **27**, pp.534-553.

[5] Hansen, P. C. (1992) *SIAM Review*, **34**, pp. 561-580.

[6] Hansen, P. C. (1992), *Regularization Tools. A Matlab Package for Analysis and Solution of Discrete Ill-Posed Problems, Manual and Tutorial*, Report UNIC-92-03, UNI•C.

[7] Hansen, P. C. and O'Leary, D. P. (1993) *SIAM J. Sci. Comput.*, **14**, pp. 1487-1503.

[8] Hanson, R. J. (1971) *SIAM J. Numer. Anal.*, **8**, pp. 616-622.

[9] Lawson, C. L. and Hanson, R. J. (1974) *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs.

[10] Phillips, D. L. (1962) *J. Assoc. Comput. Mach.*, **9**, pp. 84-97.

[11] Rust, B. W. (1998) *Truncating the Singular Value Decomposition for Ill-Posed Problems*, NISTIR 6131, National Institute of Standards and Technology.

[12] Wahba, G. (1977) *SIAM J. Numer. Anal.*, **14**, pp. 651-667.