3.7 Nonlinear Problems

The least-squares solutions examined thus far treat the coefficient matrix $\mathbf{E}$ as known. But in many of the cases encountered in practice, the elements of $\mathbf{E}$ are computed from data and are imperfectly specified. It is well known in the regression literature that treating $\mathbf{E}$ as known, even if $\mathbf{n}$ is increased beyond the errors contained in $\mathbf{y}$, can lead to significant bias errors in the least-squares and related solutions, particularly if $\mathbf{E}$ is nearly singular [91] The problem is known as that of "errors in regressors or errors in variables" (EIV); it manifests itself in the classical simple least-squares problem, P. 43, where a straight line is being fit to data of the form $y_i = a + bt_i$, but where the measurement positions, t_i, are partly uncertain rather than perfect.

In general terms, when $\mathbf{E}$ has errors, the model statement becomes

$$\left(\tilde{\mathbf{E}} + \Delta\tilde{\mathbf{E}}\right)\tilde{\mathbf{x}} = \tilde{\mathbf{y}} + \Delta\tilde{\mathbf{y}} \tag{3.56}$$ {eq:55001}

where one seeks estimates,$\tilde{\mathbf{x}}$, $\Delta\tilde{\mathbf{E}}$, $\Delta\tilde{\mathbf{y}}$ where the old $\mathbf{n}$ is now broken into two parts: $\Delta\tilde{\mathbf{E}}\tilde{\mathbf{x}}$ and $-\Delta\tilde{\mathbf{y}}$. If such estimates can be made, the result can be used to rewrite (3.56) as,

$$\tilde{\mathbf{E}}\tilde{\mathbf{x}} = \tilde{\mathbf{y}}, \tag{3.57}$$ {eq:55002}

where the relation is to be exact. That is, one seeks to modify the elements of $\mathbf{E}$ such that the observational noise in it is reduced to zero.

3.7.1 Total Least Squares

For some problems of this form, the method of total least squares (TLS) is a powerful and interesting method. It is worth examining briefly to understand why it is not always immediately useful, and to motivate a different approach.[92]

The SVD plays a crucial role in TLS. Consider, for example, that in Equation 2.17, we wrote the vector $\mathbf{y}$ as a sum of the column vectors of $\mathbf{E}$; to the extent that the column space does not fully describe $\mathbf{y}$, a residual must be left by the solution $\tilde{\mathbf{x}}$, and ordinary least squares can be regarded as producing a solution in which a new estimate, $\tilde{\mathbf{y}} \equiv \mathbf{E}\tilde{\mathbf{x}}$, of $\mathbf{y}$ is made; $\mathbf{y}$ is changed,

but the elements of $\mathbf{E}$ are untouched. But suppose it were possible to introduce small changes in both the column vectors of $\mathbf{E}$, as well as in $\mathbf{y}$, such that the column vectors of the modified $\mathbf{E} + \Delta\mathbf{E}$ produced a spanning vector space for $\mathbf{y} + \Delta\mathbf{y}$, where both $\|\Delta\mathbf{y}\|$, $\|\Delta\mathbf{x}\|$ were "small," then the problem as stated would be solved.

The simplest problem to analyze is the full-rank, formally overdetermined one. Let $M \geq N = K$. Then, if we form the $M \times (N+1)$ augmented matrix

$$\mathbf{E}_a = \{\mathbf{E} \quad \mathbf{y}\},$$

the solution we seek is such that

{eq:55003}
$$\{\tilde{\mathbf{E}} \quad \tilde{\mathbf{y}}\} \begin{bmatrix} \tilde{\mathbf{x}} \\ \\ -1 \end{bmatrix} = \mathbf{0} \tag{3.58}$$

(exactly). If this solution is to exist, $[\tilde{\mathbf{x}}, -1]^T$ must lie in the nullspace of $\{\tilde{\mathbf{E}} \quad \tilde{\mathbf{y}}\}$. A solution is thus ensured by forming the SVD of $\{\mathbf{E} \quad \mathbf{y}\}$, setting $\lambda_{N+1} = 0$, and forming $\{\tilde{\mathbf{E}} \quad \tilde{\mathbf{y}}\}$ out of the remaining singular vectors and values. Then $[\tilde{\mathbf{x}} \quad -1]^T$ is the nullspace of the modified augmented matrix, and must therefore be proportional to the nullspace vector $\mathbf{v}_{N+1}$. Also,

{eq:55004}
$$\{\Delta\tilde{\mathbf{E}} \quad \Delta\tilde{\mathbf{y}}\} = -\mathbf{u}_{N+1}\lambda_{N+1}\mathbf{v}_{N+1}^T. \tag{3.59}$$

Various complications can be considered, for example, if the last element of $\mathbf{v}_{N+1} = 0$; this and other special cases are discussed in the reference. Cases of nonuniqueness are treated by selecting the solution of minimum norm. A simple generalization applies to the underdetermined case: If the rank of the augmented matrix is p, one reduces the rank by one to $p - 1$.

The TLS solution just summarized applies only to the case in which the errors in the elements of $\mathbf{E}$ and $\mathbf{y}$ are uncorrelated and of equal variance and in which there are no required structures—for example, where certain elements of $\mathbf{E}$ must always vanish. More generally, changes in some elements of $\mathbf{E}$ require, for reasons of physics, specific corresponding changes in other elements of $\mathbf{E}$ and in $\mathbf{y}$, and vice versa. The fundamental difficulty is that the model, Eq. (3.56), presents a nonlinear estimation problem with correlated variables, and its solution requires modification of the linear procedures we have been using.

3.7.2 Method of Total Inversion

The simplest form of TLS does not readily permit the use of correlations and prior variances in the parameters appearing in the coefficient matrix and does not provide any way of maintaining the zero structure there. Methods exist that permit accounting for prior knowledge of

covariances.[93] Suppose we have a set of nonlinear constraints in a vector of unknowns $\mathbf{x}$,

{eq:55005}
$$\mathbf{g}(\mathbf{x}) + \mathbf{u} = \mathbf{q}\,. \tag{3.60}$$

This set of equations is the generalization of the linear models hitherto used; $\mathbf{u}$ again represents any expected error in the specification of the model. An example of a scalar nonlinear model is

$$8x_1^2 + x_2^2 + u = 4\,.$$

In general, there will be some expectations about the behavior of $\mathbf{u}$. Without loss of generality, we take its expected value to be zero, and its covariance is $\mathbf{Q} = \langle \mathbf{u}\mathbf{u}^T \rangle$. There is nothing to prevent us from combining $\mathbf{x}$, $\mathbf{u}$ into one single set of unknowns $\boldsymbol{\xi}$, and indeed if the model has some unknown parameters, $\boldsymbol{\xi}$ might as well include those as well. So (3.60) can be written

{eq:55006}
$$\mathcal{L}\,(\xi) = \mathbf{0} \tag{3.61}$$

In addition, it is supposed that a reasonable initial estimate $\tilde{\boldsymbol{\xi}}(0)$ is available, with uncertainty $\mathbf{P}(0) \equiv \langle \big(\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}(0)\big)\big(\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}(0)\big)^T \rangle$ (or the covariances of the $\mathbf{u}$, $\mathbf{x}$ could be specified separately if their uncertainties are not correlated). An objective function is written

{eq:55007}
$$J = \mathcal{L}(\boldsymbol{\xi})^T \mathbf{Q}^{-1} \mathcal{L}(\boldsymbol{\xi}) + \big(\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}(0)\big)\ \mathbf{P}(0)^{-1} \big(\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}(0)\big)\,, \tag{3.62}$$

whose minimum is sought. The presence of the weight matrices $\mathbf{Q}$, $\mathbf{P}(0)$ permits control of the elements most likely to change, which should not change at all [e.g., by introducing real zeros into $\mathbf{P}(0)$], as well as the stipulation of covariances. We can regard it as a generalization of the process of minimizing objective functions, which led us to least squares in previous chapters. It is sometimes known as the "method of total inversion.[94]

Consider an example for the two simultaneous equations,

{eq:55009}
$$2x_1 + x_2 + n_1 = 1 \tag{3.63}$$

{eq:55010}
$$0 + 3x_2 + n_2 = 2 \tag{3.64}$$

where all the numerical values except the zero are now regarded as in error to some degree. One way to proceed is to write the coefficients of $\mathbf{E}$ in the specific perturbation form (3.56). For example, we might write, $E_{11} = 2 + \Delta E_{11}$, and define the unknowns $\boldsymbol{\xi}$ in terms of the ΔE_{ij}. Let us for illustration retain the full nonlinear form by setting

$$\xi_1 = E_{11}, \quad \xi_2 = E_{12}, \quad \xi_3 = E_{21}, \quad \xi_4 = E_{22}, \quad \xi_5 = x_1, \quad \xi_6 = x_2,$$
$$u_1 = n_1, \quad u_2 = n_2\,.$$

The equations are then,

$$\xi_1\xi_5 + \xi_2\xi_6 + u_1 - 1 = 0 \tag{3.65}$$ {eq:55011}

$$\xi_3\xi_5 + \xi_4\xi_6 + u_2 - 2 = 0\,. \tag{3.66}$$ {eq:55012}

The y_i are being treated as formally fixed, but the presence of u_1, u_2 represent their possible errors (the division into different elements of knowns and unknowns is not unique). Let there be an initial estimate,

$$\xi_1 = 2 \pm 1, \quad \xi_2 = 2 \pm 2, \quad \xi_3 = 0 \pm 0,$$
$$\xi_4 = 3.5 \pm 1, \quad \xi_5 = x_1 = 0 \pm 2, \quad \xi_6 = 0 \pm 2\,,$$

with no imposed correlations so that $\mathbf{P}(0) = \mathrm{diag}\,([1,1,0,1,4,4])$; the zero represents the requirement that E_{21} remain unchanged. Let $\mathbf{Q} = \mathrm{diag}\,([2,2])$. Then a useful objective function is,

$$\begin{aligned} J = {} & \big(\xi_1\xi_5 + \xi_2\xi_6 - 1\big)^2/2 \\ & + \big(\xi_3\xi_5 + \xi_4\xi_6 - 2\big)^2/2 + (\xi_1 - 2)^2 + (\xi_2 - 2)^2/4 \\ & + 10^6\xi_3^2 + (\xi_4 - 3.5)^2 + \xi_5^2/4 + \xi_6^2/4\,. \end{aligned} \tag{3.67}$$ {totinv1}

The 10^6 in front of the term in ξ_3^2 is a numerical approximation to the infinite value implied by a zero uncertainty in this term (an arbitrarily large value can cause numerical instability, characteristic of penalty and barrier methods).[95]

Such objective functions define surfaces in spaces of the dimension of $\boldsymbol{\xi}$. Most procedures require the investigator to make a first guess at the solution, $\tilde{\boldsymbol{\xi}}(0)$, and attempt to minimize J by going downhill from the guess. Various search algorithms have been developed and are variants of steepest descent, conjugate gradient, Newton and quasi-Newton methods. The difficulties are numerous: Some methods require computation or provision of the gradients of J with respect to $\boldsymbol{\xi}$, and the computational cost may become very great. The surfaces on which one is seeking to go downhill may become extremely tortuous, or very slowly changing. One can fall into local holes that are not the true minima. Finding one's way is something of an art. Nonetheless, existing techniques are very useful. The minimum of J corresponds to finding the solution of the nonlinear normal equations that would result from setting the partial derivatives to zero.

Let the true minimum be at $\boldsymbol{\xi}^*$. Assuming that the search procedure has succeeded, the objective function is locally

$$J = \text{constant} + \big(\boldsymbol{\xi} - \boldsymbol{\xi}^*\big)^T\mathcal{H}\big(\boldsymbol{\xi} - \boldsymbol{\xi}^*\big) + \Delta J \tag{3.68}$$ {eq:55014}

where $\mathcal{H}$ is the Hessian and ΔJ is a correction–assumed to be small. In the linear least-squares problem, Eq. (2.89), the Hessian is evidently $\mathbf{E}^T\mathbf{E}$, the second derivative of the objective function

with respect to $\mathbf{x}$. The supposition is then that near the true optimum, the objective function is locally quadratic with a small correction. To the extent that this supposition is true, we can analyze the result in terms of the behavior of $\mathcal{H}$ as though it represented a locally defined version of $\mathbf{E}^T\mathbf{E}$. In particular, if $\mathcal{H}$ has a nullspace, or small eigenvalues, one can expect to see all the issues arising that we dealt with in Chapter 2, including ill-conditioning and solution variances that may become large in some elements. The machinery used in Chapter 2 (row and column scaling, nullspace suppression, etc.) thus becomes immediately relevant here and can be used to help conduct the search and to understand the solution.

Example 18 *It remains to find the minimum of J in (3.67).*[96] *Most investigators are best-advised to tackle problems such by using one of the many general purpose numerical routines written by experts* [97] *Here, a quasi-Newton method was employed to produce,*

$$E_{11} = 2.0001, \quad E_{12} = 1.987, \quad E_{21} = 0.0,$$
$$E_{22} = 3.5237, \quad x_1 = -0.0461, \quad x_2 = 0.556$$

and the minimum of $J = 0.0802$. The inverse Hessian at the minimum is,

$$\mathcal{H}^{-1} = \left\{ \begin{array}{cccccc} 0.4990 & 0.0082 & -0.0000 & -0.0014 & 0.0061 & 0.0005 \\ 0.0082 & 1.9237 & 0.0000 & 0.0017 & -0.4611 & -0.0075 \\ -0.0000 & 0.0000 & 0.0000 & -0.0000 & -0.0000 & 0.0000 \\ -0.0014 & 0.0017 & -0.0000 & 0.4923 & 0.0623 & -0.0739 \\ 0.0061 & -0.4611 & -0.0000 & 0.0623 & 0.3582 & -0.0379 \\ 0.0005 & -0.0075 & 0.0000 & -0.0739 & -0.0379 & 0.0490 \end{array} \right\}.$$

The eigenvalues and eigenvectors of $\mathcal{H}$ *are*

$$\lambda_i = \begin{bmatrix} 2.075 \times 10^6 & 30.4899 & 4.5086 & 2.0033 & 1.9252 & 0.4859 \end{bmatrix},$$

$$\mathbf{V} = \begin{Bmatrix} 0.0000 & -0.0032 & 0.0288 & 0.9993 & 0.0213 & 0.0041 \\ -0.0000 & 0.0381 & -0.2504 & 0.0020 & 0.0683 & 0.9650 \\ -1.0000 & 0.0000 & 0.0000 & -0.0000 & -0.0000 & 0.0000 \\ 0.0000 & 0.1382 & 0.2459 & -0.0271 & 0.9590 & -0.0095 \\ -0.0000 & 0.1416 & -0.9295 & 0.0237 & 0.2160 & -0.2621 \\ 0.0000 & 0.9795 & 0.1095 & 0.0035 & -0.1691 & 0.0017 \end{Bmatrix}.$$

The large jump from the first eigenvalue to the others is a reflection of the conditioning problem introduced by having one element, ξ_3*, with almost zero uncertainty, characteristic of barrier methods. It is left to the reader to use this information about* $\mathcal{H}$ *to compute the uncertainty of the solution in the neighborhood of the optimal values–this would be the new uncertainty,* $\mathbf{P}(1)$*. A local resolution analysis follows from that of the SVD, employing knowledge of the* $\mathbf{V}$*. The particular system is too small for a proper statistical test of the result against the prior covariances, but the possibility should be clear. If* $\mathbf{P}(0)$ *etc., are simply regarded as nonstatistical weights, we are free to experiment with different values until a pleasing solution is found.*

3.7.3 Variant Nonlinear Methods, Including Combinatorial Ones

As with the linear least-squares problems discussed in Chapter 2, many possibilities exist for objective functions that are nonlinear in either data constraint terms or the model, and there are many variations on methods for searching for objective function minima.

As with any mathematical subject dealing with nonlinearities, there are no fully general, guaranteed methods that prevent difficulties. But a very interesting and useful set of methods has been developed comparatively recently, called "combinatorial optimization." Combinatorial methods do not promise that the true minimum is found—merely that it is highly probable—because they search the space of solutions in clever ways which make it unlikely that one is very far from the true optimal solution. Two such methods, simulated annealing and genetic algorithms, have recently attracted considerable attention.[98] Simulated annealing searches randomly for solutions that reduce the objective function from a present best value. Its clever addition to purely random guessing is a willingness to accept the occasional uphill solution—one that actually raises the value of the objective function—as a way of avoiding being trapped

in purely local minima. The probability of accepting an uphill value and the size of the tried random perturbations depend upon a parameter, a temperature defined in analogy to the real temperature of a slowly cooling (annealing) solid.

Genetic algorithms, as their name would suggest, are based upon searches generated in analogy to genetic drift in biological organisms.[99] The recent literature is large and sophisticated, and this approach is not pursued here.

Notes

[60]Bracewell (1978), Freeman (1965), Jerri (1977), or Butzer and Stens (1992)

[61]In the Russian literature, Kotel'nikov's theorem.

[62]Aliasing is familiar as the stroboscope effect. Recall the appearance of the spokes of a wagon wheel in the movies. The spokes can appear to stand still, or move slowly forward or backward, depending upon the camera shutter speed relative to the true rate at which the spokes revolve. (The terminology is apparently due to John Tukey.)

[63]Hamming (1973) and Bracewell (1978) have particularly clear discussions.

[64]There is a story, perhaps apocryphal, that a group of investigators was measuring the mass flux of the Gulf Stream at a fixed time each day. They were preparing to publish the exciting discovery that there was a strong 14-day periodicity to the flow, before someone pointed out that they were aliasing the tidal currents of period 12.42 hours.

[65]It follows from the so-called Paley-Wiener criterion, and is usually stated in the form that "timelimited signals cannot be bandlimited"

[66]Landau & Pollack (1962); Freeman (1965); Jerri (1977).

[67]Petersen & Middleton (1962). An application, with discussion of the noise sensitivity, may be found in Wunsch, 1989.)

[68]Davis & Polonsky, 1965

[69]See Ripley (1981, §5.2)

[70]Bretherton *et al.* (1976).

[71]A fuller discussion may be found in Thièbaux and Pedder (1987) and Daley (1991).

[72]See Fukumori et al. (1991)

[73]Luenberger (1969).

[74]See Lawson and Hanson (1974) or Strang (1986); the standard full treatment is Fiacco and McCormick (1968).

[75]Lawson and Hanson (1974).

[76]Fu (1981).

[77]Tziperman and Hecht (1987).

[78]Dantzig (1963).

[79]For example, Luenberger (1984); Bradley, Hax, & Magnanti (1977); and many others.

[80]See Strang, 1986; Luenberger, 1969; Cacuci, 1981; Hall & Cacuci, 1984; Rockafellor, 1993.

[81]One of a few mathematical algorithms ever to be written up on the front page of The New York Times (19 November 1984, story by J. Gleick)–a reflection of the huge economic importance of linear programs in industry.

[82]Arthnari & Dodge (1981).

[83]Wagner (1969); Arthnari & Dodge (1981)

[84]Van Huffel & Vandewalle (1991)

[85]The use of EOFs, with various normalizations, scalings, and in various row/column physical spaces, is widespread—for example, Wallace and Dickinson (1972), Wallace (1972), Davis (1978b), and many others.

[86]Jolliffe (1986); Preisendorfer (1988); Jackson (1991)

[87]Davenport & Root (1958); Wahba (1990)

[88]Berkooz, Holmes, & Lumley (1993),

[89]Armstrong (1989), David (1988), Ripley (1981)

[90]Ripley (1981).

[91]For example, Seber (1977).

[92]Golub and van Loan (1980, 1989) and Van Huffel and Vandewalle (1991).

[93]Tarantola and Valette (1982) and Tarantola (1987).

[94]Tarantola and Valette (1982) labeled the use of similar objective functions and the determination of the minimum as the *method of total inversion*, although they considered only the case of perfect model constraints.

[95]Luenberger (1984

[96]Tarantola and Valette (1982) suggested using a linearized search method, iterating from the initial estimate, which must be reasonably close to the correct answer. The method can be quite effective (e.g., Wunsch & Minster, 1982; Mercier, 1986). In a wider context, however, their method is readily recognizable as a special case of the many known methods for minimizing a general objective function.

[97]Numerical Algorithms Group (1988); Grace (1990; Press et al. (1992)..

[98]For simulated annealing, the literature starts with Pincus (1970) and Kirkpatrick, Gelatt, and Vecchi (1983), and general discussions can be found in van Laarhoven and Aarts (1987), Ripley (1981), and Press et al. (1992). A simple oceanographic application to experiment design was discussed by Barth and Wunsch (1989).

[99]Goldberg (1989), Holland (1992), and Denning (1992).