12.3 Other Examples of the Least-Squares Principle

The principle of least squares is also used in other situations. In one of these, we attempt to *solve* an inconsistent system of linear equations of the form

$$\sum_{j=0}^{n} a_{kj} x_j = b_k \qquad (0 \leqq k \leqq m) \tag{1}$$

in which $m > n$. Here, there are $m + 1$ equations but only $n + 1$ unknowns. If a given $n+1$-tuple $(x_0, x_1, \ldots, x_n)$ is substituted on the left, the discrepancy between the two sides of the kth equation is termed the ***k*th residual**. Ideally, of course, all residuals should be zero. If it is not possible to select $(x_0, x_1, \ldots, x_n)$ so as to make all residuals zero, System (1) is said to be **inconsistent** or **incompatible**. In this case, an alternative is to minimize the sum of the squares of the residuals. So we are led to minimize the expression

$$\varphi(x_0, x_1, \ldots, x_n) = \sum_{k=0}^{m} \left(\sum_{j=0}^{n} a_{kj} x_j - b_k \right)^2 \tag{2}$$

by making an appropriate choice of $(x_0, x_1, \ldots, x_n)$. Proceeding as before, we take partial derivatives with respect to x_i and set them equal to zero, thereby arriving at the normal equations

$$\sum_{j=0}^{n} \left(\sum_{k=0}^{m} a_{ki} a_{kj} \right) x_j = \sum_{k=0}^{m} b_k a_{ki} \qquad (0 \leqq i \leqq n) \tag{3}$$

This is a linear system of just $n + 1$ equations involving unknowns $x_0, x_1, \ldots, x_n$. It can be shown that this system is consistent, provided that the column vectors in the original coefficient array are linearly independent. System (3) can be solved, for instance, by Gaussian elimination. The solution of System (3) is then a best approximate solution of Equation (1) in the least-squares sense.

Special methods have been devised for the problem just discussed. Generally, they gain in precision over the simple approach outlined above. One such algorithm for solving System (1),

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

begins by factoring

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$$

where matrix $\boldsymbol{Q}$ is $(m+1) \times (n+1)$ satisfying $\boldsymbol{Q}^T \boldsymbol{Q} = \boldsymbol{I}$ and matrix $\boldsymbol{R}$ is $(n+1) \times (n+1)$ satisfying $r_{ii} > 0$ and $r_{ij} = 0$ for $j < i$. Then the least-squares solution is obtained by an algorithm called the *modified Gram-Schmidt process*.

A more elaborate (and more versatile) algorithm depends on the **singular value decomposition** of the matrix $\boldsymbol{A}$. This is a factoring, $\boldsymbol{A} = \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, in which $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_{m+1}$, $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_{n+1}$, and $\boldsymbol{\Sigma}$ is an $(m + 1) \times (n + 1)$ diagonal matrix that has nonnegative entries. For these more reliable procedures, the reader is referred to material at the end of this section and to Stewart [1973] and Lawson and Hanson [1995].

Use of a Weight Function $w(x)$

Another important example of the principle of least squares occurs in fitting or approximating functions on *intervals* rather than discrete sets. For example, a given function f defined on an interval $[a, b]$ may have to be approximated by a function such as

$$g(x) = \sum_{j=0}^{n} c_j g_j(x)$$

It is natural, then, to attempt to minimize the expression

$$\varphi(c_0, c_1, \ldots, c_n) = \int_a^b [g(x) - f(x)]^2 \, dx \tag{4}$$

by choosing coefficients appropriately. In some applications, it is desirable to force functions g and f into better agreement in certain parts of the interval. For this purpose, we can modify Equation (4) by including a positive **weight function** $w(x)$, which can, of course, be $w(x) \equiv 1$ if all parts of the interval are to be treated the same. The result is

$$\varphi(c_0, c_1, \ldots, c_n) = \int_a^b [g(x) - f(x)]^2 w(x) \, dx$$

The minimum of φ is again sought by differentiating with respect to each c_i and setting the partial derivatives equal to zero. The result is a system of normal equations:

$$\sum_{j=0}^{n} \left[\int_a^b g_i(x) g_j(x) w(x) \, dx \right] c_j = \int_a^b f(x) g_i(x) w(x) \, dx \qquad (0 \leqq i \leqq n) \tag{5}$$

This is a system of $n+1$ linear equations in $n+1$ unknowns $c_0, c_1, \ldots, c_n$ and can be solved by Gaussian elimination. Earlier remarks about choosing a good basis apply here also. The ideal situation is to have functions $g_0, g_1, \ldots, g_n$ that have the orthogonality property:

$$\int_a^b g_i(x) g_j(x) w(x) \, dx = 0 \qquad (i \neq j) \tag{6}$$

Many such orthogonal systems have been developed over the years. For example, **Chebyshev polynomials** form one such system, namely,

$$\int_{-1}^{1} T_i(x) T_j(x) (1 - x^2)^{-1/2} \, dx = \begin{cases} 0 & i \neq j \\ \dfrac{\pi}{2} & i = j > 0 \\ \pi & i = j = 0 \end{cases}$$

The weight function $(1 - x^2)^{-1/2}$ assigns heavy weight to the ends of the interval $[-1, 1]$.

If a sequence of nonzero functions $g_0, g_1, \ldots, g_n$ is orthogonal according to Equation (6), then the sequence $\lambda_0 g_0, \lambda_1 g_1, \ldots, \lambda_n g_n$ is orthonormal for appropriate positive real numbers λ_j, namely,

$$\lambda_j = \left\{ \int_a^b [g_j(x)]^2 w(x) \, dx \right\}^{-1/2}$$

Nonlinear Example

As another example of the least-squares principle, here is a nonlinear problem. Suppose that a table of points (x_k, y_k) is to be fitted by a function of the form

$$y = e^{cx}$$

Proceeding as before leads to the problem of minimizing the function

$$\varphi(c) = \sum_{k=0}^{m} (e^{cx_k} - y_k)^2$$

The minimum occurs for a value of c such that

$$0 = \frac{\partial \varphi}{\partial c} = \sum_{k=0}^{m} 2(e^{cx_k} - y_k)e^{cx_k}x_k$$

This equation is nonlinear in c. One could contemplate solving it by Newton's method or the secant method. On the other hand, the problem of minimizing $\varphi(c)$ could be attacked directly. Since there can be multiple roots in the normal equation and local minima in φ itself, a direct minimization of φ would be safer. This type of difficulty is typical of **nonlinear least-squares problems**. Consequently, other methods of curve fitting are often preferred if the unknown parameters do not occur linearly in the problem.

Alternatively, this particular example can be linearized by a change of variables $z = \ln y$ and by considering

$$z = cx$$

The problem of minimizing the function

$$\varphi(c) = \sum_{k=0}^{m} (cx_k - z_k)^2 \qquad z_k = \ln y_k$$

is easy and leads to

$$c = \sum_{k=0}^{m} z_k x_k \Big/ \sum_{k=0}^{m} x_k^2$$

This value of c is *not* the solution of the original problem but may be satisfactory in some applications.

Linear and Nonlinear Example

The final example contains elements of linear and nonlinear theory. Suppose that an (x_k, y_k) table is given with $m + 1$ entries and that a functional relationship such as

$$y = a \sin(bx)$$

is suspected. *Can the least-squares principle be used to obtain the appropriate values of the parameters a and b?*

Notice that parameter b enters this function in a nonlinear way, creating some difficulty, as will be seen. According to the principle of least squares, the parameters should be chosen such that the expression

$$\sum_{k=0}^{m} [a \sin(bx_k) - y_k]^2$$

has a minimum value. The minimum value is sought by differentiating this expression with respect to a and b and setting these partial derivatives equal to zero. The results are

$$\begin{cases} \displaystyle\sum_{k=0}^{m} 2[a \sin(bx_k) - y_k] \sin(bx_k) = 0 \\ \displaystyle\sum_{k=0}^{m} 2[a \sin(bx_k) - y_k] a x_k \cos(bx_k) = 0 \end{cases}$$

If b were known, a could be obtained from either equation. The correct value of b is the one for which these corresponding two a values are identical. So each of the preceding equations should be solved for a, and the results set equal to each other. This process leads to the equation

$$\frac{\sum_{k=0}^{m} y_k \sin bx_k}{\sum_{k=0}^{m} (\sin bx_k)^2} = \frac{\sum_{k=0}^{m} x_k y_k \cos bx_k}{\sum_{k=0}^{m} x_k \sin bx_k \cos bx_k}$$

which can now be solved for parameter b, using, for example, the bisection method or the secant method. Then either side of this equation can be evaluated as the value of a.

Additional Details on SVD

The singular value decomposition (SVD) of a matrix is a factorization that can reveal important properties of the matrix that otherwise could escape detection. For example, from the SVD decomposition of a square matrix one could be alerted to the near-singularity of the matrix. Or from the SVD factorization of a nonsquare matrix an unexpected loss of rank could be revealed. Since the SVD factorization of a matrix yields a complete orthogonal decomposition, it provides a technique for computing the least squares solution of a system of equations and at the same time producing the norm of the error vector.

Suppose that a given $m \times n$ matrix has the factorization

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m]$ is an $m \times m$ orthogonal matrix, $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$ is an $n \times n$ orthogonal matrix, and the $m \times n$ diagonal matrix $\boldsymbol{D}$ contains the singular values of $\boldsymbol{A}$ on its diagonal, listed in decreasing order. The singular values of a matrix $\boldsymbol{A}$ are the positive square roots of the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$. These are denoted by $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$. In detail, we have

$$\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V} = \boldsymbol{D} = \begin{bmatrix} \sigma_1 & & & & & & \\ & \sigma_2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_r & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}_{m \times n}$$

where $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_m$ and $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_n$. (In the above matrix, blank space corresponds to zero entries.) Moreover, we have $\boldsymbol{A}\boldsymbol{v}_i = \sigma_i\boldsymbol{u}_i$ and $\sigma_i = ||\boldsymbol{A}\boldsymbol{v}_i||_2$ where $\boldsymbol{v}_i$ is column i in $\boldsymbol{V}$ and

$\boldsymbol{u}_i$ is column i in $\boldsymbol{U}$. Since $\boldsymbol{U}$ is orthogonal, we obtain

$$\begin{aligned}
||\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}||_2^2 &= ||\boldsymbol{U}^T(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})||_2^2 = ||\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{x}-\boldsymbol{U}^T\boldsymbol{b}||_2^2 \\
&= ||\boldsymbol{U}^T\boldsymbol{A}(\boldsymbol{V}\boldsymbol{V}^T)\boldsymbol{x}-\boldsymbol{U}^T\boldsymbol{b}||_2^2 \\
&= ||(\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V})(\boldsymbol{V}^T\boldsymbol{x})-\boldsymbol{U}^T\boldsymbol{b}||_2^2 \\
&= ||\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{x}-\boldsymbol{U}^T\boldsymbol{b}||_2^2 = ||\boldsymbol{D}\boldsymbol{y}-\boldsymbol{c}||_2^2 \\
&= \sum_{i=1}^{r}(\sigma_i y_i - c_i)^2 + \sum_{i=r+1}^{m} c_i^2
\end{aligned}$$

where $\boldsymbol{y}=\boldsymbol{V}^T\boldsymbol{x}$ and $\boldsymbol{c}=\boldsymbol{U}^T\boldsymbol{b}$. Here, $\boldsymbol{y}$ is defined by $y_i = c_i/\sigma_j$ and $\boldsymbol{x}$ by $\boldsymbol{x}=\boldsymbol{V}\boldsymbol{y}$. Since $c_i=\boldsymbol{u}_i^T\boldsymbol{b}$ and $\boldsymbol{x}=\boldsymbol{V}\boldsymbol{y}$, if $y_i=\sigma_i^{-1}c_i$ for $1 \leqq i \leqq r$ then the least-squares solution is

$$\boldsymbol{x}_{LS} = \sum_{i=1}^{n} y_i\boldsymbol{v}_i = \sum_{i=1}^{r}\sigma_i^{-1}c_i\boldsymbol{v}_i = \sum_{i=1}^{r}\sigma_i^{-1}\left(\boldsymbol{u}_i^T\boldsymbol{b}\right)\boldsymbol{v}_i$$

and

$$||\boldsymbol{A}\boldsymbol{x}_{LS}-\boldsymbol{b}||_2^2 = \sum_{i=r+1}^{m} c_i^2 = \sum_{i=r+1}^{m}\left(\boldsymbol{u}_i^T\boldsymbol{b}\right)^2$$

which is the smallest of all two-norm minimizers. For additional, details see Golub and Van Loan [1996].

In conclusion, we obtain the following theorem.

■ **THEOREM 1**

SVD LEAST SQUARES THEOREM

Let $\boldsymbol{A}$ be an $m \times n$ matrix of rank r. Let the SVD factorization be $\boldsymbol{A}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$. The least-squares solution of the system $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ is $\boldsymbol{x}_{LS}=\sum_{i=1}^{n}(\sigma_i^{-1}c_i)\boldsymbol{v}_i$, where $c_i=\boldsymbol{u}_i^T\boldsymbol{b}$. If there exist many least-squares solutions to the given system, then the one of least 2-norm is $\boldsymbol{x}$ as described above.

EXAMPLE 1 Find the least-squares solution of this nonsquare system

$$\begin{bmatrix}1 & 1\\ 0 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}x\\ y\\ z\end{bmatrix} = \begin{bmatrix}1\\ -1\\ 1\end{bmatrix}$$

using the singular value decomposition:

$$\begin{bmatrix}1 & 1\\ 0 & 1\\ 1 & 0\end{bmatrix} = \begin{bmatrix}\frac{1}{3}\sqrt{6} & 0 & \frac{1}{3}\sqrt{3}\\ \frac{1}{6}\sqrt{6} & \frac{1}{2}\sqrt{2} & -\frac{1}{3}\sqrt{3}\\ \frac{1}{6}\sqrt{6} & -\frac{1}{2}\sqrt{2} & -\frac{1}{3}\sqrt{3}\end{bmatrix}\begin{bmatrix}\sqrt{3} & 0\\ 0 & 1\\ 0 & 0\end{bmatrix}\begin{bmatrix}\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2}\\ \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2}\end{bmatrix}$$

Solution We have $r=\text{rank}(\boldsymbol{A})=2$ and the singular values $\sigma_1=\sqrt{3}$ and $\sigma_2=1$. This leads to

$$c_1=\boldsymbol{u}_1^T\boldsymbol{b} = \begin{bmatrix}\frac{1}{3}\sqrt{6} & \frac{1}{6}\sqrt{6} & \frac{1}{6}\sqrt{6}\end{bmatrix}\begin{bmatrix}1\\ -1\\ 1\end{bmatrix} = \frac{1}{3}\sqrt{6}$$

and

$$c_2 = \boldsymbol{u}_2^T \boldsymbol{b} = \begin{bmatrix} 0 & -\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = \sqrt{2}$$

and

$$x_{LS} = (\sigma_1^{-1} c_1)\boldsymbol{v}_1 + (\sigma_2^{-1} c_2)\boldsymbol{v}_2 = \frac{1}{\sqrt{3}} \left(\frac{1}{3}\sqrt{6}\right) \begin{bmatrix} \frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \end{bmatrix} + \sqrt{2} \begin{bmatrix} \frac{1}{2}\sqrt{2} \\ -\frac{1}{2}\sqrt{2} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{4}{3} \\ -\frac{2}{3} \end{bmatrix}$$

This solution is the same as that from the normal equations. ■

Using the Singular Value Decomposition

This material requires the theory of the singular value decomposition discussed in Section 8.3.

An important application of the singular value decomposition is in the matrix least-squares problem, to which we now return. For any system of linear equations $\boldsymbol{Ax} = \boldsymbol{b}$, we want to define a unique **minimal solution**. This is described as follows. Let $\boldsymbol{A}$ be $m \times n$, and define

$$\rho = \inf\{||\boldsymbol{Ax} - \boldsymbol{b}||_2 : \boldsymbol{x} \in \mathbb{R}^n\}$$

The minimal solution of our system is taken to be the point of smallest norm in the set $\{\boldsymbol{x}: ||\boldsymbol{Ax} - \boldsymbol{b}||_2 = \rho\}$. If the system is *consistent*, then $\rho = 0$, and we are simply asking for the point of least norm among all solutions. If the system is *inconsistent*, we want $\boldsymbol{Ax}$ to be as close as possible to $\boldsymbol{b}$; that is, $||\boldsymbol{Ax} - \boldsymbol{b}||_2 = \rho$. If there are many such points, we choose the one closest to the origin.

The minimal solution is produced by using the *pseudo-inverse* of $\boldsymbol{A}$, and this object, in turn, can be computed from the singular value decomposition of $\boldsymbol{A}$ as discussed in Section 8.3. First, consider a diagonal $m \times n$ matrix of the following form, where the σ_j are positive numbers:

$$\boldsymbol{D} = \begin{bmatrix} \sigma_1 & & & & & & \\ & \sigma_2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_r & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}_{m \times n}$$

Its pseudo-inverse $\boldsymbol{D}^+$ is defined to be of the same form, except that it is to be $n \times m$ and it has $1/\sigma_j$ on its diagonal. For example,

$$\boldsymbol{D} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix} \qquad \boldsymbol{D}^+ = \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{2} \\ 0 & 0 \end{bmatrix}$$

If $\boldsymbol{A}$ is any $m \times n$ matrix and if $\boldsymbol{UDV}^T$ is one of its **singular value decompositions**, we define the **pseudo-inverse** of $\boldsymbol{A}$ to be

$$\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{D}^+\boldsymbol{U}^T$$

We do not stop to prove that the pseudo-inverse of $\boldsymbol{A}$ is unique if we impose the order $\sigma_1 \geqq \sigma_2 \geqq \cdots$.

■ **THEOREM 2**

MINIMAL SOLUTION THEOREM

Consider a system of linear equations $\boldsymbol{Ax} = \boldsymbol{b}$, in which $\boldsymbol{A}$ is an $m \times n$ matrix. The minimal solution of the system is $\boldsymbol{A}^+\boldsymbol{b}$.

Proof Use the notation established above, and let $\boldsymbol{x}$ be any point in $\mathbb{R}^n$. Define $\boldsymbol{y} = \boldsymbol{V}^T\boldsymbol{x}$ and $\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{b}$. Using the properties of $\boldsymbol{V}$ and $\boldsymbol{U}$, we obtain

$$\begin{aligned} \rho &= \inf_{\boldsymbol{x}} ||\boldsymbol{Ax} - \boldsymbol{b}||_2 \\ &= \inf_{\boldsymbol{x}} ||\boldsymbol{UDV}^T\boldsymbol{x} - \boldsymbol{b}||_2 \\ &= \inf_{\boldsymbol{x}} ||\boldsymbol{U}^T(\boldsymbol{UDV}^T\boldsymbol{x} - \boldsymbol{b})||_2 \\ &= \inf_{\boldsymbol{x}} ||\boldsymbol{DV}^T\boldsymbol{x} - \boldsymbol{U}^T\boldsymbol{b}||_2 \\ &= \inf_{\boldsymbol{y}} ||\boldsymbol{Dy} - \boldsymbol{c}||_2 \end{aligned}$$

Exploiting the special nature of $\boldsymbol{D}$, we have

$$\left|\left|\boldsymbol{Dy} - \boldsymbol{c}\right|\right|_2^2 = \sum_{i=1}^{r}(\sigma_i y_i - c_i)^2 + \sum_{i=r+1}^{m} c_i^2$$

To minimize this last expression, we define $y_i = c_i/\sigma_i$ for $1 \leqq i \leqq r$. The other components can remain unspecified. But to get the $\boldsymbol{y}$ of least norm, we must set $y_i = 0$ for $r+1 \leqq i \leqq m$. This construction is carried out by the pseudo-inverse $\boldsymbol{D}^+$, so $\boldsymbol{y} = \boldsymbol{D}^+\boldsymbol{c}$. Hence, we obtain

$$\boldsymbol{x} = \boldsymbol{Vy} = \boldsymbol{VD}^+\boldsymbol{c} = \boldsymbol{VD}^+\boldsymbol{U}^T\boldsymbol{b} = \boldsymbol{A}^+\boldsymbol{b}$$

Let us express the minimal solution in another form, taking advantage of the zero components in the vector $\boldsymbol{y}$. Since $y_i = 0$ for $i > r$, we require only the first r components of $\boldsymbol{y}$. These are given by $y_i = c_i/\sigma_i$. Now it is evident that only the first r components of $\boldsymbol{c}$ are needed. Since $\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{b}$, c_i is the inner product of row i in $\boldsymbol{U}^T$ with the vector $\boldsymbol{b}$. That is the same as the inner product of the ith column of $\boldsymbol{U}$ with $\boldsymbol{b}$. Thus,

$$\boldsymbol{y}_i = \boldsymbol{u}_i^T\boldsymbol{b}/\sigma_i \qquad 1 \leqq i \leqq r$$

The minimal solution, which we may denote by $\boldsymbol{x}^*$, is then

$$\boldsymbol{x}^* = \boldsymbol{V}\boldsymbol{y} = \sum_{i=1}^{r} y_i \boldsymbol{v}_i \qquad \blacksquare$$

An example of this procedure can be carried out in mathematical software such as Matlab, Maple or Mathematica. We can generate a system of 20 equations with three unknowns by a random process. This technique is often used in testing software, especially in benchmarking studies, in which a large number of examples is run with careful timing. The software has a provision for entering random matrices. When executed, the computer program first exhibits the random input. The three singular values of matrix $\boldsymbol{A}$ are displayed. Then the diagonal 20×3 matrix $\boldsymbol{D}$ is displayed. A check on the numerical work is made by computing $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, which should equal $\boldsymbol{A}$. Then the pseudo-inverse of $\boldsymbol{D}^+$ is computed. Next, the pseudo-inverse $\boldsymbol{A}^+$ is computed. The minimal solution, $\boldsymbol{x} = \boldsymbol{A}^+\boldsymbol{b}$, is computed, as well as the residual vector, $\boldsymbol{r} = \boldsymbol{A}^+\boldsymbol{b} = \boldsymbol{b}$. Then the orthogonality condition $\boldsymbol{A}^T\boldsymbol{r} = \boldsymbol{0}$ is checked. This program is therefore carrying out all the steps described above for obtaining the minimal solution of a system of equations. Another example will be given below to show what happens in the case of a loss in rank. (See Computer Problem 12.3.10.)

In problems of this type, the user must examine the singular values and decide whether any are small enough to warrant being set equal to zero. The necessity of this step becomes clear when we look at the definition of $\boldsymbol{D}^+$. The reciprocals of the singular values are the principal constituents of this matrix. Any very small singular value that is *not* set equal to zero will therefore have a disruptive effect on the subsequent calculations. A rule of thumb that has been recommended is to drop any singular value whose magnitude is less than σ_1 times the inherent accuracy of the coefficient matrix. Thus, if the data are accurate to three decimal places and if $\sigma_1 = 5$, then any σ_i less than 0.005 should be set equal to zero.

An example of a small matrix having a **near-deficiency in rank** is given next. In the Maple program, certain singular values are set equal to zero if they fail to meet the relative size criterion mentioned in the previous paragraph. Also, we have added, as a check on the calculations, a verification of the following four **Penrose properties** for a pseudo-matrix.

■ **THEOREM 3** **PENROSE PROPERTIES OF THE PSEUDO-INVERSE**

The pseudo-inverse $\boldsymbol{A}^+$ for the matrix $\boldsymbol{A}$ has these four properties:

$$\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} \qquad \boldsymbol{A}^+ = \boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+$$
$$\boldsymbol{A}\boldsymbol{A}^+ = (\boldsymbol{A}\boldsymbol{A}^+)^T \qquad \boldsymbol{A}^+\boldsymbol{A} = (\boldsymbol{A}^+\boldsymbol{A})^T$$

We can use mathematical software such as Matlab, Maple, or Mathematica for finding the pseudo-inverse of a matrix that has a deficiency in rank. For example, consider this 5×3 matrix:

$$\boldsymbol{A} = \begin{bmatrix} -85 & -55 & -115 \\ -35 & 97 & -167 \\ 79 & 56 & 102 \\ 63 & 57 & 69 \\ 45 & -8 & 97.5 \end{bmatrix} \tag{7}$$

A tolerance value is set so that in the evaluation of singular values any value whose magnitude is less than the tolerance is treated as zero. We can verify the Penrose properties for this matrix. (See Computer Problem 12.3.11.)

Summary

(1) We attempt to solve an **inconsistent system**

$$\sum_{j=0}^{n} a_{kj}x_j = b_k \qquad (0 \leqq k \leqq m)$$

in which there are $m+1$ equations but only $n+1$ unknowns with $m > n$. We minimize the sum of the squares of the residuals and are led to minimize the expression

$$\varphi(x_0, x_1, \ldots, x_n) = \sum_{k=0}^{m} \left(\sum_{j=0}^{n} a_{kj}x_j - b_k \right)^2$$

We solve the $(n+1) \times (n+1)$ system of normal equations

$$\sum_{j=0}^{n} \left(\sum_{k=0}^{m} a_{ki}a_{kj} \right) x_j = \sum_{k=0}^{m} b_k a_{ki} \qquad (0 \leqq i \leqq n)$$

by Gaussian elimination, and the solution is a best approximate solution of the original system in the least-squares sense.

Additional References

See Acton [1959], Björck [1996], Branham [1990], Cheney [1982, 2001], Forsythe [1957], van Huffel and Vandewalle [1991], Lawson and Hanson [1995], Rice [1971], Rice and White [1964], Rivlin [1990], Späth [1992], and Whittaker and Robinson [1944].