Notes on systems of equations with no solution

Measuring errors

For systems of equations $A\boldsymbol{x} = \boldsymbol{b}$ the following basic theorem is known characterizing when there is a solution $\boldsymbol{x}$.

Exactly one of the following holds:
(I) $A\boldsymbol{x} = \boldsymbol{b}$ has a solution $\boldsymbol{x}$ *or* (II) $\boldsymbol{y}A = \boldsymbol{0}\boldsymbol{y}\boldsymbol{b} \neq 0$ has a solution $\boldsymbol{y}$

That is, either the system has a solution or it is inconsistent in an 'obvious' way.

However, in some situations when there is no solution we may want to find a 'best' approximation to an exact solution. Consider the following simple example

$$\begin{array}{rcrcr} x_1 & - & 4x_2 & = & -5 \\ -x_1 & + & x_2 & = & 2 \\ 2x_1 & + & 3x_2 & = & 7 \\ -2x_i & - & 5x_2 & = & -9 \end{array}$$

which has no solution. If we try, for example $x_1 = 2, x_2 = 1$ the left side of the first equation is $2-4(1) = -2$. The error is the difference between the right side and the left side. In this case the error is $-5-(-2) = -3$. Similarly the error for the second equation evaluated at $x_1 = 2, x_2 = 1$ is $e_2 = 2-(-2+1) = 3$ and the errors for the other equations are $e_3 = 7-(2(2)+3(1)) = 0$ and $e_4 = -9-(-2(2)-5) = 0$. So the error vector for $(2,1)$ is $(e_1, e_2, e_3, e_4) = (-3, 3, 0, 0)$. If we try another point $x_1 = 2, x_2 = 1$ we get an error vector $(2, 1, -1, 3)$ and if we try $x_1 = 1, x_2 = 1$ we get error vector $(-2, 2, 2, -2)$. Which of these is 'best'? (Note that there might be better fractional solutions here but we will stick to these three proposed solutions for illustration.)

What 'best' is depends on our measure of the size of the error vectors
$(-3, 3, 0, 0), (2, 1, -1, 3), (-2, 2, 2, -2)$. If we minimize the sum of the absolute values of the errors the first is best, if we minimize the sum of the squares of the errors the second is best and if we minimize the largest absolute value of an error we get the third.

Slightly more formally, the L_p norm of a vector $(e_1, e_2, \ldots, e_n)$ is $(\sum_{i=1}^{n} |e_i|^p)^{1/p}$. Minimizing the sum of the absolute values of the errors minimizes the L_1 norm. Minimizing the sum of the squares of the errors minimizes the L_2 norm and minimizing the largest absolute value minimizes the L_∞ norm.

It is convenient to express the system of equations using $\sum$ notation. Thus we will look at systems of m equations in the n variables $x_1, x_2, \ldots, x_n$ written as follows:

$\sum_{j=1}^{n} a_{ij}x_j = b_i$ for $i = 1, 2, \ldots, m$. We will also use the matrix notation for this, $A\boldsymbol{x} = \boldsymbol{b}$.

The errors for a given $x_1^*, x_2^*, \ldots, x_n^*$ are $e_i = b_i - \sum_{j=1}^{n} a_{ij}x_j$. So we get the following expressions:

The best L_1 approximating solution(s) are x_i minimizing $\sum_{i=1}^{m} |b_i - \sum_{j=1}^{n} a_{ij}x_j|$.

The best L_2 approximating solution(s) are x_i minimizing $\sum_{i=1}^{m} (b_i - \sum_{j=1}^{n} a_{ij}x_j)^2$.

The best L_∞ approximating solution(s) are x_i minimizing $\max_i |b_i - \sum_{j=1}^{n} a_{ij}x_j|$.

We will see the we can find the best L_1 and L_∞ approximating solutions by solving a linear programming problem and the best L_2 approximating solution (least squares solution) by doing some elementary calculus and then solving a system of equations.

Best L_1 and L_∞ approximations

In order to deal with the absolute values we recall that $|a|$ is a if a is nonnegative and $-a$ is a is negative. So a variable x that satisfies $x \geq a$ and $x \geq -a$ satisfies $x \geq |a|$. Thus we can set up variables that are at least as large as the absolute values of the errors and when we minimize we will in fact get values equal to the absolute values.

For the L_1 norm we introduce a new variable f_i for each error $|b_i - \sum_{j=1}^n a_{ij}x_j|$. We constrain $f_i \geq b_i - \sum_{j=1}^n a_{ij}x_j$ and $f_i \geq -(b_i - \sum_{j=1}^n a_{ij}x_j)$. Then minimizing $\sum_{i=1}^n f_i$ in the following linear program, for each possible choice of the x_i, each of the f_i will be equal to the corresponding error and the x_i in an optimal solution are the x_i for a best approximating L_1 solution. The inequalities in the previous sentence are rearranged so that the variables are all on the left side.

$$\begin{array}{lrcl}
\min & \sum_{i=1}^n 0x_i \;+\; \sum_{i=1}^n f_i & & \\
\\
\text{s.t.} & (\sum_{j=1}^n a_{ij}x_j) \;+\; f_i & \geq & b_i \quad \text{for } i = 1, 2, \ldots, m \\
& -(\sum_{j=1}^n a_{ij}x_j) \;+\; f_i & \geq & -b_i \quad \text{for } i = 1, 2, \ldots, m
\end{array}$$

Note that in the i^{th} inequality the coefficient for f_k, $k \neq i$ is 0.

For the example above we get

$$\begin{array}{rrcrcrcrcrcrl}
\min & 0x_1 & + & 0x_2 & + & f_1 & + & f_2 & + & f_3 & + & f_4 & \\
\text{s.t} & x_1 & - & 4x_2 & + & f_1 & & & & & & & \geq -5 \\
& -x_1 & + & x_2 & & & + & f_2 & & & & & \geq 2 \\
& 2x_1 & + & 3x_2 & & & & & + & f_3 & & & \geq 7 \\
& -2x_1 & - & 5x_2 & & & & & & & + & f_4 & \geq -9 \\
& -x_1 & + & 4x_2 & + & f_1 & & & & & & & \geq 5 \\
& x_1 & - & x_2 & & & + & f_2 & & & & & \geq -2 \\
& -2x_1 & - & 3x_2 & & & & & + & f_3 & & & \geq -7 \\
& 2x_1 & + & 5x_2 & & & & & & & + & f_4 & \geq 9
\end{array}$$

For the L_∞ norm we introduce a single variable z for all of the errors $|b_i - \sum_{j=1}^n a_{ij}x_j|$. We constrain $z \geq b_i - \sum_{j=1}^n a_{ij}x_j$ and $z \geq -(b_i - \sum_{j=1}^n a_{ij}x_j)$. Then minimizing z in the following linear program, for each possible choice of the x_i, will make z equal to the largest of the absolute values and the x_i in an optimal solution are the x_i for a best approximating L_∞ solution. The inequalities in the previous sentence are rearranged so that the variables are all on the left side.

$$\begin{array}{lrcl}
\min & \sum_{i=1}^n 0x_i \;+\; z & & \\
\\
\text{s.t.} & (\sum_{j=1}^n a_{ij}x_j) \;+\; z & \geq & b_i \quad \text{for } i = 1, 2, \ldots, m \\
& -(\sum_{j=1}^n a_{ij}x_j) \;+\; z & \geq & -b_i \quad \text{for } i = 1, 2, \ldots, m
\end{array}$$

For the example above we get

$$\begin{array}{rrcrcrcr} \min & 0x_1 & + & 0x_2 & + & z & & \\ \text{s.t} & x_1 & - & 4x_2 & + & z & \geq & -5 \\ & -x_1 & + & x_2 & + & z & \geq & 2 \\ & 2x_1 & + & 3x_2 & + & z & \geq & 7 \\ & -2x_1 & - & 5x_2 & + & z & \geq & -9 \\ & -x_1 & + & 4x_2 & + & z & \geq & 5 \\ & x_1 & - & x_2 & + & z & \geq & -2 \\ & -2x_1 & - & 3x_2 & + & z & \geq & -7 \\ & 2x_1 & + & 5x_2 & + & z & \geq & 9 \end{array}$$

Best L_2 approximation

For the best L_2 approximation we use basic calculus. For each $k = 1, 2, \ldots, n$ consider the partial derivative $\frac{\partial}{\partial x_k}\left(\sum_{i=1}^m (b_i - \sum_{j=1}^n a_{ij}x_j)^2\right) = \sum_{i=1}^m 2\left(b_i - \sum_{j=1}^n a_{ij}x_j\right)(-a_{ik}x_k)$. Setting these equal to 0 we have, for $k = 1, 2 \ldots, n$ (after a little algebra to rearrange terms), $\sum_{j=1}^n (\sum_{i=1}^n a_{ik}a_{ij})x_j = \sum_{i=1}^m b_i a_{ik}$. With a little work one can check that this becomes the following in matrix notation $A^TA\boldsymbol{x} = A^T\boldsymbol{b}$. These are called the normal equations. We note that under reasonable conditions on the columns of A (they are linearly independent) A^TA will have an inverse and the equations give the best solution as $(A^TA)^{-1}A^T\boldsymbol{b}$.

In the example above we have $A^TA = \begin{bmatrix} 1 & -1 & 2 & -2 \\ -4 & 1 & 3 & -5 \end{bmatrix} \begin{bmatrix} 1 & -4 \\ -1 & 1 \\ 2 & 3 \\ -2 & -5 \end{bmatrix} = \begin{bmatrix} 10 & 11 \\ 11 & 51 \end{bmatrix}$ and $A^T\boldsymbol{b} = \begin{bmatrix} 1 & -1 & 2 & -2 \\ -4 & 1 & 3 & -5 \end{bmatrix} \begin{bmatrix} -5 \\ 2 \\ 7 \\ -9 \end{bmatrix} = \begin{bmatrix} 25 \\ 88 \end{bmatrix}$. So $A^TA\boldsymbol{x} = \boldsymbol{b}$ is $\begin{bmatrix} 10 & 11 \\ 11 & 51 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 25 \\ 88 \end{bmatrix}$.

This is the system of equations $\begin{array}{rcrcr} 10x_1 & + & 11x_2 & = & 25 \\ 11x_1 & + & 51x_2 & = & 88 \end{array}$ with solution $x_1 = 307/389$ and $x_2 = 605/389$.

Geometrically this also tell us something. If we take A times the least squares solution we get $\begin{bmatrix} 1 & -4 \\ -1 & 1 \\ 2 & 3 \\ -2 & -5 \end{bmatrix} \begin{bmatrix} 307/389 \\ 605/389 \end{bmatrix} = \frac{1}{389} \begin{bmatrix} -2113 \\ 298 \\ 2429 \\ -3639 \end{bmatrix}$. This is the projection of $\boldsymbol{b} = \begin{bmatrix} -5 \\ 2 \\ 7 \\ -9 \end{bmatrix}$ onto the column space of A. (This the space spanned by the columns of A. For example if we had only 3 rows and 2 columns then the column space would be the unique plane in three dimensional space containing the vectors given by the 2 columns of A.)