3.3 Inequality Constraints; Nonnegative Least Squares

In many estimation problems, it is useful to be able to impose inequality constraints upon the solutions. Problems involving tracer concentrations, for example, usually demand that they remain positive; empirical eddy diffusion coefficients are sometimes regarded as acceptable only when non-negative; in some fluid flow problems we may wish to impose directions, but not magnitudes, upon velocity fields.

Such needs lead to consideration of the forms,

{eq:51002}
$$\mathbf{Ex} + \mathbf{n} = \mathbf{y}, \tag{3.35}$$

{eq:51003}
$$\mathbf{Gx} \geq \mathbf{h}\,, \tag{3.36}$$

Continued on next page...

where the use of a greater-than inequality to represent the general case is purely arbitrary; multiplication by minus 1 readily reversing it. $\mathbf{G}$ is of dimension $M_2 \times N$.

Several cases need to be distinguished. (A) Suppose $\mathbf{E}$ is full rank and fully determined; then the SVD solution to (3.35) by itself is $\tilde{\mathbf{x}}$, $\tilde{\mathbf{n}}$, and there is no solution nullspace. Substitution of the solution into (3.36) shows that the inequalities are either satisfied or that some are violated. In the first instance, we are finished, and the inequalities bring no new information. In the second case, the solution must be modified and necessarily, $\|\tilde{\mathbf{n}}\|$ will increase, given the noise-minimizing nature of the SVD solution. It is also possible that the inequalities are contradictory, in which case there is no solution.

(B) Suppose that $\mathbf{E}$ is formally underdetermined—so that a solution nullspace exists. If the particular-SVD solution violates one or more of the inequalities and requires modification, we can distinguish two subcases. (1) Addition of one or more nullspace vectors permits the inequalities to be satisfied. Then the solution residual norm will be unaffected, but $\|\tilde{\mathbf{x}}\|$ will increase. (2) The nullspace vectors by themselves are unable to satisfy the inequality constraints, and one or more range vectors are required to do so. Then both $\|\tilde{\mathbf{x}}\|$, $\|\tilde{\mathbf{n}}\|$ will increase.

Case (A) is the conventional one.[74] The so-called Kuhn-Tucker-Karush theorem is a requirement for a solution $\tilde{\mathbf{x}}$ to exist. Its gist is as follows: Let $M \geq N$ and $\mathbf{E}$ be full rank; there are no $\mathbf{v}_i$ in the solution nullspace. If there is a solution, there must exist a vector, $\mathbf{q}$, of dimension M_2, such that

$$\mathbf{E}^T(\mathbf{E}\tilde{\mathbf{x}} - \mathbf{y}) = \mathbf{G}^T\mathbf{q}\,. \tag{3.37}$$ {eq:51004}

$$\mathbf{G}\mathbf{x} - \mathbf{h} = \mathbf{r}\,, \tag{3.38}$$ {eq:51005}

where the M_2 elements of $\mathbf{q}$ are divided into two groups. For group 1, of dimension m_1,

$$r_i = 0, \qquad q_i \geq 0\,, \tag{3.39}$$ {eq:51006}

and for group 2, of dimension $m_2 = M_2 - m_1$,

$$r_i > 0, \qquad q_i = 0\,. \tag{3.40}$$ {eq:51007}

To understand this theorem, recall that in the solution to the ordinary overdetermined least-squares problem, the left-hand side of (3.37) vanishes identically (2.91 and 2.262), being the projection of the residuals onto the range vectors, $\mathbf{u}_i$, of $\mathbf{E}^T$. If this solution violates one or more of the inequality constraints, one must introduce into it structures that produce increased residuals.

Because there are no nullspace $\mathbf{v}_i$, the rows of $\mathbf{G}$ may each be expressed exactly by an expansion in the range vectors. In the second group of indices, the corresponding inequality constraints are already satisfied by the ordinary least-squares solution, and no modification of the structure proportional to $\mathbf{v}_i$ is required. In the first group of indices, the inequality constraints are marginally satisfied, at equality, only by permitting violation of the demand (2.91) that the residuals should be orthogonal to the range vectors of $\mathbf{E}$. If the ordinary least-squares solution violates the inequality, the minimum modification required to it pushes the solution to the edge of the acceptable bound, but at the price of increasing the residuals proportional to the corresponding $\mathbf{u}_i$. The algorithm consists of finding the two sets of indices and then the smallest coefficients of the $\mathbf{v}_i$ corresponding to the group 1 indices required to just satisfy any initially violated inequality constraints. A canonical special case, to which more general problems can be reduced, is based upon the solution to $\mathbf{G} = \mathbf{I}$, $\mathbf{h} = \mathbf{0}$—called "nonnegative least squares".[75] The requirement, $\mathbf{x} \geq 0$, is essential in many problems involving tracer concentrations, which are neccessarily positive.

The algorithm can be extended to the underdetermined/rank-deficient case in which the addition, to the original basic SVD solution, of appropriate amounts of the nullspace of $\mathbf{v}_i$ is capable of satisfying any violated inequality constraints.[76] One simply chooses the smallest mean-square solution coefficients necessary to push the solution to the edge of the acceptable inequalities, producing the smallest norm. The residuals of the original problem do not increase—because only nullspace vectors are being used. $\mathbf{G}$ must have a special structure for this to be possible.

The algorithm can be further generalized[77] by considering the general case of rank-deficiency/-underdeterminism where the nullspace vectors by themselves are inadequate to produce a solution satisfying the inequalities. In effect, any inequalities "left over" are satisfied by invoking the smallest perturbations necessary to the coefficients of the range vectors $\mathbf{v}_i$.