Stanford Exploration Project, Report 94, November 11, 1997, pages 1–141

Short Note

Preconditioning and scaling

Jon Claerbout [1]

keywords: *adjoint, imaging, inversion, least-squares, regression, weighting*

INTRODUCTION

In geophysical mapping and imaging applications we set up linear equations of high order. We face subjective issues like how to scale components of an operator and how much damping (regularization) to use. Here I summarize a few scaling tricks. Please keep in mind that we do not have matrices (data structures) but operators, i.e., function pairs for applying an operator $\mathbf{F}$ and its adjoint (transpose). In other words, nifty methods and theories for matrices are often not applicable.

ROW NORMALIZATION, POSITIVE OPERATORS

Perhaps most of the operators in Geophysics are of the special type whose associated matrix contains no negative elements. Examples are stacking, migration, velocity analysis, DMO, tomography, integration, and linear interpolation. For such positive operators $\mathbf{F}$ it can be informative to apply an input vector that has components that are all ones. Denote such a vector by $\mathbf{1}$. The vector $\mathbf{F1}$ tells us the sum of the elements on any row of the associated matrix. We can define a diagonal matrix by spreading the vector $\mathbf{F1}$ on the diagonal of a matrix. Denote this by $\mathbf{diag}(\mathbf{F1})$. We can smooth this diagonal and add a small threshhold $\epsilon > 0$ so as to avoid any problem with inversion. Such a diagonal scaling operator might be denoted as $\mathbf{D} = \mathbf{diag}(< \mathbf{F1} > +\epsilon)$. In various applications we might find it more useful to work with the row normalized operator $\mathbf{D}^{-1}\mathbf{F}$ than with the operator $\mathbf{F}$ itself.

ROW NORMALIZATION, NON-POSITIVE OPERATORS

Sometimes an operator $\mathbf{D}$ does not contain all positive numbers, but we can easily form another operator which, element by element, has the absolute values of the

[1] **email:** jon@geo.stanford.edu

elements of $\mathbf{D}$. An example of this arises with estimation of prediction-error filters where the matrix $\mathbf{D}$, contains raw seismic data. It is easy enough to replace each seismic data value by its absolute value. I'll write this as $|\mathbf{D}|$ where it is understood that the absolute values are applied to each component in the matrix $\mathbf{D}$. In reality we would not have the matrix itself, but we would make a filter of all ones, and convolve it with the seismic magnitudes; I'll write this as $|\mathbf{D}|\mathbf{1}$. Obviously this absolute-valued data could be smoothed; it could have a tiny threshold ϵ added; and it could then be inverted and used as a weighting function.

A use of this weighting function is when we are estimating a prediction-error filter from noisy data. We choose a scaling factor for each fitting equation equal to the inverse to the sum of the absolute values of the elements on the row. Thus, where there are spikes in the data, we have scaled each defective fitting equation by the inverse of the spike (the spike plus other positive values). Notice that we still have a linear problem. Since we are not scaling by the inverse residual, the solution is still easy weighted least squares; we don't need complicated robust L_1 solvers.

ROUGHENERS AND SMOOTHERS

I remind you of the "standard trick" used by Bill Symes, Bill Harlan, and probably many others. In geophysical fitting we generally have two goals, the first being data "fitting" and the second being the "damping" or "regularization" goal.

$$
\begin{array}{rcccl}
0 & \approx & \mathbf{r}_d & = & \mathbf{F}\mathbf{m} - \mathbf{d} \qquad (1) \\
0 & \approx & \mathbf{r}_m & = & \lambda \mathbf{R}\mathbf{m} \qquad (2)
\end{array}
$$

where $\mathbf{R}$ is a roughening operator. Since the roughening operator $\mathbf{R}$ is somewhat arbitrarily chosen, presumably we can know an inverse to it, a smoothing operator $\mathbf{S}$ so that $\mathbf{RS} = \mathbf{I}$. In the simple one-dimensional case, the roughener is a first or second derivative operator and the smoother is a single or double integral. (I'm not so sure that the multidimensional case is as easy as we often like to pretend). We define a new variable $\mathbf{x}$ by $\mathbf{m} = \mathbf{S}\mathbf{x}$ and have a set of regression goals for which iterative methods generally converge faster:

$$
\begin{array}{rcccl}
0 & \approx & \mathbf{r}_d & = & \mathbf{F}\mathbf{S}\mathbf{x} - \mathbf{d} \qquad (3) \\
0 & \approx & \mathbf{r}_m & = & \lambda \mathbf{x} \qquad (4)
\end{array}
$$

After solving for $\mathbf{x}$ we easily find $\mathbf{m} = \mathbf{S}\mathbf{x}$.

A problem arises when we do not start from $\mathbf{m} = 0$. Then before we can get started we need to find $\mathbf{x}_0$ by solving $\mathbf{m}_0 = \mathbf{S}\mathbf{x}_0$.

HOW MUCH DAMPING?

We often need to damp the solutions to least squares problems. We have a fitting problem (regression) with the two goals:

$$
\begin{aligned}
0 &\approx \mathbf{r}_d = \mathbf{Fm} - \mathbf{d} && (5)\\
0 &\approx \mathbf{r}_m = \lambda \mathbf{Rm} && (6)
\end{aligned}
$$

where $\mathbf{R}$ is a roughening operator. How big should λ be? Suppose in human terms we'd like "half" the properties of the solution to come from the fitting and "half" to come from the damping. How might we define what we mean by "half"? We can start by considering balancing the two residual vectors $\mathbf{r}_d$ and $\mathbf{r}_m$.

$$
\lambda_0 = \frac{||\mathbf{Fm} - \mathbf{d}||}{||\mathbf{Rm}||} \tag{7}
$$

Another approach is to balance the gradients. The gradient is the "force" on the solution m.

$$
\lambda_1 = \frac{||\mathbf{F}'(\mathbf{Fm} - \mathbf{d})||}{||\mathbf{R}'\mathbf{Rm}||} \tag{8}
$$

where $\mathbf{F}'$ is the transpose of $\mathbf{F}$ and likewise for $\mathbf{R}'$. I suspect that λ_1 is the better choice for λ but a little more experience would add confidence.

ROW PARTITIONED OPERATORS

A problem arises with partitioned operators. Here we are fitting observed data to theoretical data where there are two classes of model parameters $\mathbf{m}_1$ and $\mathbf{m}_2$. We seek to minimize the norm of the residual $\mathbf{r}$ defined by

$$
\mathbf{0} \approx \mathbf{r} = [\alpha\mathbf{F} \quad \beta\mathbf{B}] \begin{bmatrix} \mathbf{m}_1/\alpha \\ \mathbf{m}_2/\beta \end{bmatrix} - \mathbf{d} \tag{9}
$$

where α and β are arbitrary scaling constants. The residual is independent of α and β, but when we "solve" this system using (as we must) the idea that $\mathbf{F}^{-1} \approx \mathbf{F}'$, we see the result depends on α and β, namely it contains α^2 and β^2

$$
\begin{bmatrix} \hat{\mathbf{m}}_1/\alpha \\ \hat{\mathbf{m}}_2/\beta \end{bmatrix} = \begin{bmatrix} \alpha\mathbf{F}' \\ \beta\mathbf{B}' \end{bmatrix} \mathbf{d} \tag{10}
$$

Let us find the best α and β. Inserting the image (10) into the residual (9) we get

$$
\begin{aligned}
\mathbf{0} &\approx \mathbf{r} = [\alpha\mathbf{F} \quad \beta\mathbf{B}] \begin{bmatrix} \alpha\mathbf{F}' \\ \beta\mathbf{B}' \end{bmatrix} - \mathbf{d} && (11)\\
\mathbf{0} &\approx \mathbf{r} = \alpha^2\mathbf{FF}'\mathbf{d} + \beta^2\mathbf{BB}'\mathbf{d} - \mathbf{d} && (12)\\
\mathbf{0} &\approx \mathbf{r} = \alpha^2\mathbf{u} + \beta^2\mathbf{v} - \mathbf{d} && (13)
\end{aligned}
$$

which defines two vectors $\mathbf{u}$ and $\mathbf{v}$.

We find the best scaling factors by setting to zero the derivative of $||\mathbf{r}||^2$ with respect to α^2 and β^2

$$\mathbf{0} \quad = \quad \begin{bmatrix} \mathbf{u'u} & \mathbf{u'v} \\ \mathbf{v'u} & \mathbf{v'v} \end{bmatrix} \begin{bmatrix} \alpha^2 \\ \beta^2 \end{bmatrix} - \begin{bmatrix} \mathbf{u'd} \\ \mathbf{v'd} \end{bmatrix} \tag{14}$$

solving gives

$$\begin{bmatrix} \alpha^2 \\ \beta^2 \end{bmatrix} = \frac{1}{\mathbf{v'v}\ \mathbf{u'u} - (\mathbf{u'v})^2} \begin{bmatrix} \mathbf{v'v} & -\mathbf{u'v} \\ -\mathbf{u'v} & \mathbf{u'u} \end{bmatrix} \begin{bmatrix} \mathbf{u'd} \\ \mathbf{v'd} \end{bmatrix} \tag{15}$$

I believe it can be shown that the values α^2 and β^2 are positive.

Recalling that $\mathbf{u} = \mathbf{FF'd}$ and $\mathbf{v} = \mathbf{BB'd}$, let us now define $\mathbf{a} = \mathbf{F'd}$ and $\mathbf{b} = \mathbf{B'd}$ so $\mathbf{u} = \mathbf{Fa}$ and $\mathbf{v} = \mathbf{Bb}$.

In imaging applications we customarily ignore the scaling factor which is the common part of α and β, namely, the denominator determinant in (15). We have the proportions

$$\alpha^2 \sim (\mathbf{b'B'Bb})(\mathbf{d'FF'd}) - (\mathbf{a'F'Bb})(\mathbf{d'BB'd}) \tag{16}$$

$$\alpha^2 \sim (\mathbf{b'B'Bb})(\mathbf{a'a}) - (\mathbf{a'F'Bb})(\mathbf{b'b}) \tag{17}$$

$$\beta^2 \sim -(\mathbf{a'F'Bb})(\mathbf{a'a}) + (\mathbf{a'F'Fa})(\mathbf{b'b}) \tag{18}$$

A deeper problem of interest arises when we seek the best diagonal matrix scaling. Then we replace α^2 with two diagonal matrices, one before and one after $\mathbf{F}$. Likewise with β and $\mathbf{B}$. We need help finding those four diagonal matrices.