PARSIMONIOUS DECONVOLUTION

Jon F. Claerbout

Abstract

Conventional deconvolution produces a glut of innovation impulses, most of which represent high frequency noise. Varimax deconvolution decreases the noise. More realistic estimated filters b_τ and further noise reduction can be achieved by finding and using the filter which maximizes $\Sigma\, x_t^2 \,\ell n\, x_t^2$ subject to $\Sigma\, x_t^2 = 1$ where x_t is the filter's input and x_t convolved with b_τ is the observed filter output.

INTRODUCTION

The problem of deconvolution is this: In nature, there is an unknown sequence of innovations x_t. Various physical processes have the effect of filtering the innovations with some unknown filter b_t. The output of the filter may or may not be combined with some additive noise u_t before the final observations y_t are made. Letting $*$ denote convolution, we have

$$x_t \; * \; b_t \; + \; u_t \; = \; y_t \tag{1}$$

where everything on the left of the equal sign is unknown.

In conventional deconvolution, one presumes the existence of a filter a_t which is causal and inverse to b_t (this is the minimum phase assumption). A sketchy description of conventional estimation of a_t is to let $u_t = 0$, $a_{-t} = 0$; $a_0 = 1$, and minimize $E(a)$ where

$$E(a) \;=\; \sum_t \,(a_t * y_t)^2 \;=\; \sum_t \,(x_t)^2 \tag{2}$$

Such a procedure has a reasonable theoretical foundation when x_t is a sample of an uncorrelated Gaussian random process. The practical problem is that x_t turns out to have an innovation at every time point, and if the data y_t is sampled at a denser rate, then the deconvolved data x_t has just that many more innovations. This result is associated with a Gaussian probability function for x_t. Common efforts to make more sense out of the innovations employ either bandpass filtering or inserting a gap in the filter, i.e., $(a_1, a_3 \ldots a_k) = 0$.

Claerbout and Muir (Robust Modeling of Erratic Data, Geophysics 1973) suggested that perhaps, medians or the L_1 norm could be used to suppress this overabundance of insignificant innovations. The application to deconvolution was more fully defined in 1975 in SEP 5 - page 134. In 1976, SEP report 10 contained seven articles on non-Gaussian modeling. In the spring of 1977, Ken Larner of the Western Geophysical Company presented a "late paper" at the European Association of Exploration Geophysics meeting on the Western Geophysical "MED" process. This process was developed by Ralph A. Wiggins (now at Del Mar Technical Associates, P.O. Box 1083, Del Mar, California 92014) who provided us with a preprint entitled "Minimum Entropy Deconvolution" to appear in a Dutch Geophysical Journal.* Wiggins was the first to demonstrate a practical technique for exploiting the non-Gaussian nature of correctly deconvolved seismograms.

Subsequently, another approach to the problem, "Deconvolution with the ℓ_1 Norm" was presented at the Fall 1977 SEG Convention by Howard L. Taylor, S. C. Banks, and J. F. McCoy.

---

*Another reference to this material is: Wiggins, R.A., 1977, Minimum Entropy Deconvolution, Proceedings of the International Symposium on Computer Aided Seismic Analysis and Discrimination, June 9 & 10, 1977, Falmouth, Mass. IEEE Computer Society, pp. 7-14.

Theory and Method

Wiggins' method was to apply the varimax technique of factor analysis to the problem of deconvolution. Specifically, he maximized

$$V(a) = \sum_{\text{channels}} \frac{\sum x_t^{\,4}}{\left(\sum x_t^{\,2}\right)^2} \tag{3}$$

Let us define the n-th norm as

$$\| x \| = \left(\sum_{\text{time}} \sum_{\text{channels}} |x_t|^n \right)^{1/n} \tag{4}$$

Wiggins' criterion is similar to minimizing the ratio of the second norm to that of the fourth norm. Since varimax involves the 4-th power of the data, it is very sensitive to the largest data values, and very insensitive to smaller data values. In contrast, the philosophy of much of our work emphasized the importance of small data values. The purpose of the present study is to combine the strengths of the two approaches. Consequently, I set out to minimize

$$N_2^1(a) = \frac{\| x \|_1}{\| x \|_2} \tag{5}$$

Early efforts to minimize (5) were rather discouraging. The criterion itself seemed fine in the sense that its minimum did seem to coincide with subjectively desirable solutions. The difficulty was that the gradient of (5) did not seem to provide a good direction for descent. To help understand what was happening, consider the perturbation of some filter a_k by an amount $\alpha\ da_k$. The numerator of (5) then takes the form

$$\text{Num}(\alpha) = \sum_t \left| \sum_\tau y_{t-\tau} (a_\tau + \alpha\ da_\tau) \right| \tag{6a}$$

$$= \sum_t \left| (x_t + \alpha\ dx_t) \right| \tag{6b}$$

A graph of (6b) versus α is a piecewise linear function of α. The slope is almost constant, but it undergoes a jump discontinuity whenever there is a sign change of the argument of the absolute value function in any term of the sum. The denominator of (5) is the square root of a parabola, namely

$$\mathrm{Den}(\alpha) = \left[\sum_t (x_t + \alpha\, dx_t)^2\right]^{1/2} \tag{7}$$

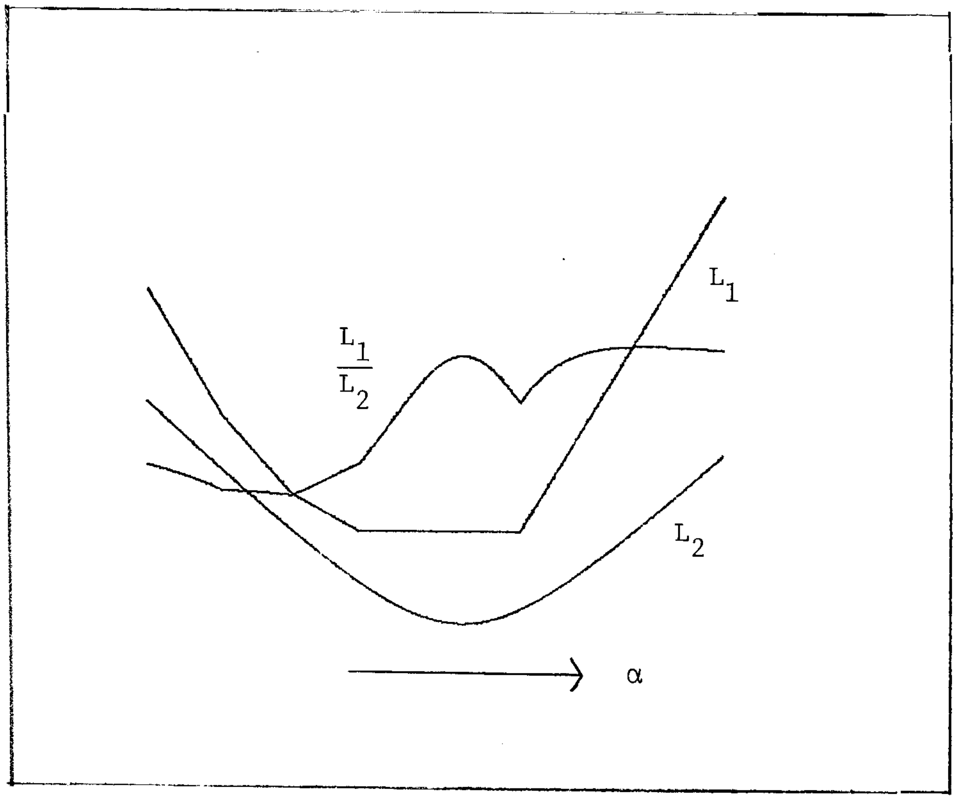


Figure 1.—Diagramatic sketch of L_1, L_2, and the ratio of L_1/L_2 for equations (5) - (7). Corners in L_1 are where absolute value functions have become zero. The ratio has multiple minima but they are easily found.

Figure 1 is a diagramatic sketch of the numerator, denominator, and ratio of the two. The possibility for many local minima is obvious. These minima would have to occur where $0 = x_t + \alpha\, dx_t$ so it is easy enough to evaluate (5) at each value $\alpha = -x_t/dx_t$ and find the global minimum. The practical problem was that the gradient of (5) with respect to a_τ did not seem to provide particularly sensible directions da_τ along which to scan α. This whole approach was abandoned before a relevant reference (*Mixed L_1 and L_2 Norm Problems*) by Luis Canales, SEP 10, p. 114) was noted.

Experimentation discovered that more sensible gradients could be obtained from

$$N_2^{3/2} = \frac{||x||_{3/2}}{||x||_2} \tag{8}$$

The idea of using the 3/2 power arose with little theoretical justification. The fact that it seemed to be a practical improvement raised the question, "what about $N_2^{5/4}$, $N_2^{7/4}$, or $N_2^{2-\varepsilon}$ where ε tends to zero?" The third possibility was particularly intriguing but it requires some analysis before it can be put to practical use. The discussion will continue after some algebra.

Let us formally consider

$$S''_n = \lim_{\varepsilon \to 0} \frac{||x||_{n-\varepsilon}}{||x||_n} \tag{9}$$

For computational purposes, we will want to eliminate the ε by analytical means. Since minimizing a function is equivalent to minimizing its logarithm, we may minimize

$$S'_n = \lim_{\varepsilon \to 0} \ln \frac{||x||_{n-\varepsilon}}{||x||_n} \tag{10a}$$

$$= \lim_{\varepsilon \to 0} \ln ||x||_{n-\varepsilon} - \ln ||x||_n \tag{10b}$$

If we expand S'_n in a power series about $\varepsilon = o$, we get

$$S'_n = 0 + \varepsilon \frac{d}{d\varepsilon} \ln ||x||_{n-\varepsilon} + 0 \ (\varepsilon^2) \tag{11}$$

Clearly, we obtain the desired limit as $\varepsilon \to 0$, if we minimize

$$S_n = \frac{d}{d\varepsilon} \ln \| x \|_{n-\varepsilon} \tag{12}$$

Insert the definition of norm (4) into (12)

$$S_n = \frac{d}{d\varepsilon} \ln \left(\Sigma |x|^{n-\varepsilon}\right)^{\frac{1}{n-\varepsilon}} \tag{13}$$

$$S_n = \frac{d}{d\varepsilon} \frac{1}{n-\varepsilon} \ln \left(\Sigma |x|^{n-\varepsilon}\right) \tag{14}$$

Before differentiating in (14), recall from fundamentals that for any constant c

$$\frac{d}{dy} c^u = c^u \ln c \frac{du}{dy}$$

so taking $y = \varepsilon$, $c = |x|$ and $u = (n-\varepsilon)$ we have

$$\frac{d}{d\varepsilon} |x|^{n-\varepsilon} = -|x|^{n-\varepsilon} \ln|x| \tag{15}$$

Now, carry through the differentiation in (14) utilizing (15)

$$S_n = \frac{\ln \Sigma |x|^{n-\varepsilon}}{(n-\varepsilon)^2} + \frac{1}{n-\varepsilon} \frac{-\Sigma |x|^{n-\varepsilon} \ln|x|}{\Sigma |x|^{n-\varepsilon}}$$

Set $\varepsilon = 0$ and rescale by the constant n^2

$$S_n = \ln \Sigma |x|^n - \frac{\Sigma |x|^n \ln|x|^n}{\Sigma |x|^n} \tag{16}$$

Note that since the value of S_n is scale invariant for x, we may introduce a constraint $\Sigma |x|^n = 1$ without affecting (16). The definition of S in (16) takes a particularly intriguing form when $n = 2$. Let $p = x^2$. Then

$$S_2 = -\Sigma\, p\, \ell n\, p \qquad \text{subject to} \quad 1 = \Sigma p \qquad (17)$$

Now, we note a formal similarity between (17) and the thermodynamic definition of entropy. Regretably, I have found it impossible to exploit this apparent similarity, primarily because p represents power, not probability.

Let us return to the question of whether better deconvolutions could be had by minimizing N_2^1, $N_2^{5/4}$, $N_2^{3/2}$, $N_2^{7/4}$, or $N_2^{2-\varepsilon}$. Computing experience seemed to indicate that all tended to be minimized for subjectively good deconvolutions. But the minimization procedure itself did not behave very reliably for any of them. Since N_2^1 had many local minima, I abandoned all but $N_2^{2-\varepsilon}$ and sought other ways to reorganize the descent procedure. It was soon discovered that convergence could be more quickly attained with $n > 2$ in S_n. But the final minimum no longer corresponded to a subjectively satisfactory deconvolution. For the examples being studied, quickest convergence was obtained by going first to a minimum with $n = 2.5$, then reducing n to 2 for final descent.

The program still behaved erratically with many unpredictable things happening. Much more comprehensible behavior was observed when the adjustable parameters were no longer taken to be the inverse filter a_t but the forward filter b_t. The forward filter is presumably a physical waveform which goes to a reasonable limit as data sampling density increases whereas the inverse filter is not so well behaved. The switch over in parameterization from the inverse filter a_t to the forward filter b_t also made it easy to explicitly enforce a causality constraint on b_t. The procedure was to permit only positive lags on b_t. The iteration would begin with all b_t equal zero except, say b_4 equal one. As the iterative descent proceeded, it almost always turned out that the maximum value on the final filter b_t turned out to be where the starting pulse was given, say b_4.

The Algorithm

The first step in the algorithm is to be able to compute what may be called the unconstrained gradient, that is, the derivative of (16) with respect to x_t. Thus, a subroutine must be available to compute g_t where

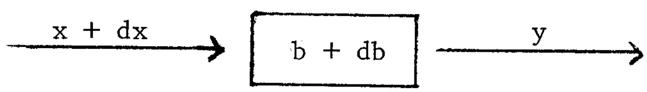
$$g_t = \frac{dS_n}{dx_t} = \frac{1}{\Sigma|x|^n}\left(1 + \frac{\Sigma|x|^n|\ell n|x|^n}{\Sigma|x|^n} - \ell n|x|^n - 1\right)\frac{d|x_t|^n}{dx_t}$$

$$= \frac{n|x|^{n-1}\ \mathrm{sgn}(x_t)}{\Sigma|x|^n}\left(\frac{\Sigma|x|^n \ell n|x|^n}{\Sigma|x|^n} - \ell n|x|^n\right) \qquad (18)$$

For $n = 2$, we have

$$g_t \cong x_t \ (\mathrm{const} - 2\ \ell n|x|) \qquad (18a)$$

Next, we recall the block diagram



The data is given to be y. Let us use X, B, Y and G to denote fourier transforms of x, b, y, and g. Clearly

$$X\ dB + B\ dX = 0 \qquad (19)$$

Now, we would like to choose dX to be some distance α in the direction of negative G. Thus,

$$X\ dB = \alpha\ G\ B \qquad (20)$$

Multiply both sides through by the conjugate of X

$$X^*\ X\ dB = \alpha\ X^*\ G\ B \qquad (21)$$

This sort of equation is ammenable to solution in the time domain for db of bounded duration but better results were obtained by using a sort of a Widrow descent procedure. What worked best was to replace the positive function $X^{*}X$ by a constant and absorb the constant into α. Then, the result could be transformed into the time domain, say

$$db \leftarrow dB = \alpha \quad X^{*} G B \tag{22}$$

Next, db can be tapered or truncated to be causal and of finite duration. It was found that sometimes db turned out to be very nearly parallel to b so that even if α was quite large, $b + \alpha$ db would be almost the same as a scaled up version of b. Rescaling b is like rescaling x and that has no effect at all on S. To avoid such deception, it makes sense to remove the projection of b on db. Namely

$$db \leftarrow db - \frac{(db,\ b)}{(b,\ b)} \, b \tag{23}$$

Unlike a_t, b_t represents a physical waveform. If a final result of 1% accuracy is desired, it seems reasonable to stop the descent procedure when db is 1% of b. But we have not yet chosen the scale factor α. The algorithmic technique was this: Decide beforehand to make some fixed large number (say 30) of iterations. Start initially with $\alpha = 20\%$ and decrease α linearly to 1% or zero.

After each updating of b by

$$b \leftarrow b + \alpha \, db \tag{24}$$

the new deconvolved data x can be found by

$$x \leftarrow X = Y/B \tag{25}$$

Actually, a more cautious approach is to use

$$X = Y \quad B^{*}/(\varepsilon + B^{*}B) \tag{26}$$

where ε bears some relation to the step size α dB or to the presumed ambient noise.